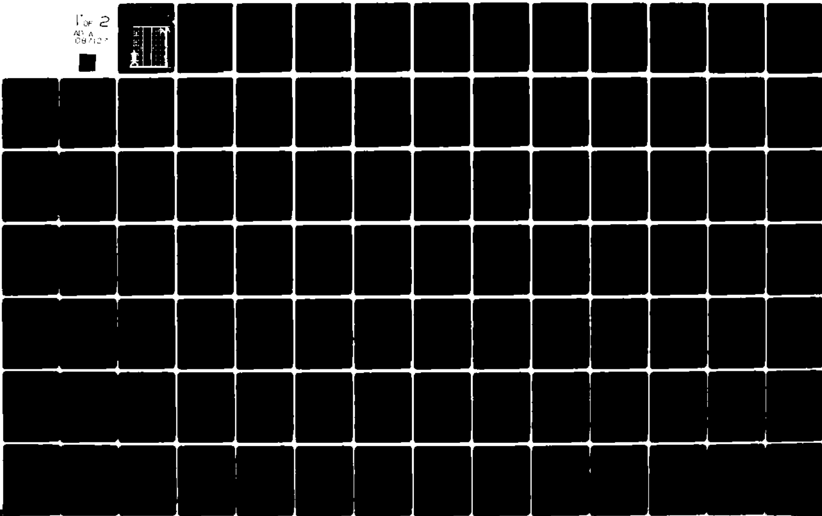
AD-A087 127 TENNESSEE UNIV KNOXVILLE DEPT OF PSYCHOLOGY F/G 5/9
RESEARCH ON THE MULTIPLE-CHOICE TEST ITEM IN JAPAN: TOWARD THE --ETC(U)
APR 80 F SAMEJIMA N00014-77-C-0360

UNCLASSIFIED ONRT-M3 NL

1 OF 2
AD A
087127



APRIL 1980

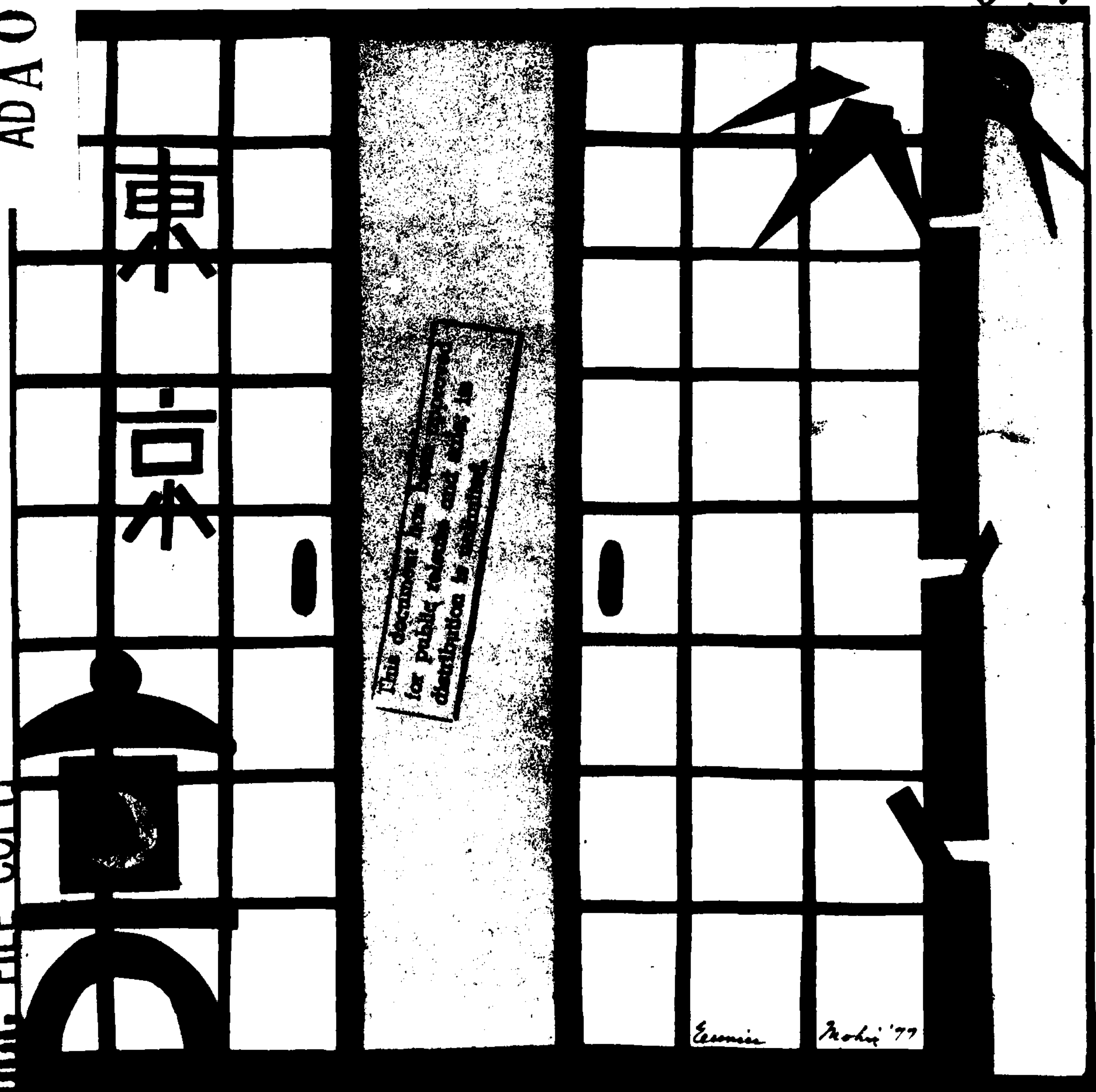
ONRT M3

SCIENTIFIC MONOGRAPH

DEPARTMENT OF THE NAVY OFFICE OF NAVAL RESEARCH TOKYO

RESEARCH ON THE MULTIPLE-CHOICE TEST ITEM IN JAPAN: TOWARD THE VALIDATION OF MATHEMATICAL MODELS

Fumiko Samejima



ONR Tokyo Scientific Monograph Series

ONRT M1 High Pressure Science and Technology in Japan by Earl F. Skelton, Naval Research Laboratory, Washington, D.C., July 1978

ONR-38 An Overview of Material Science and Engineering in Japan by George Sandoz, ONR Branch Office, Chicago, Illinois, December 1977 (Reprinted in 1980)

ONRT M2 Japanese Research Institutes Funded by The Ministry of Education, compiled by Seikoh Sakiyama, Office of Naval Research, Tokyo, January 1980

During the same period in which it was establishing ONR Tokyo and the ONR Tokyo Scientific Monograph Series, the Office of Naval Research supported and sponsored scientific liaison in Japan which led to monographs on selected segments of Japanese science. Those early monographs, which were authored in various elements of the ONR organization, are listed here:

ONR-28 Superconducting Technology in Japan by Richard G. Brandt, ONR Branch Office, Pasadena, California, June 1971

ONR-32 A Review of Some Psychological Research in Japan by Morton A. Bertin, ONR Branch Office, Chicago, Illinois, November 1972

ONR-34 Computer Science in Japan by Richard L. Lau, ONR Branch Office, Pasadena, California, June 1973

ONR-36 Chemical Science in Japan by Arnet L. Powell, ONR Branch Office, Boston, Massachusetts, March 1973.

(9) Scientific monographs

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER (14) (18) ONRT/M3

2. GOVT ACCESSION NO. AD-A087 427

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle) (6) Research on the Multiple-Choice Test item in Japan: Toward the Validation of Mathematical Models

5. TYPE OF REPORT & PERIOD COVERED (11) Apr 80

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s) (10) Dr. Fumiko Samejima
Department of Psychology
University of Tennessee
Knoxville, TN 37916

8. CONTRACT OR GRANT NUMBER(s) (15) N00014-77-C-0360

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Office of Naval Research
Scientific Liaison Group
American Embassy
APO San Francisco 96503
(12) 97

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 150-402

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
536 South Clark Street
Chicago, IL 60605

12. REPORT DATE April 1980

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
Personnel and Training Research Program
Psychological Sciences Division
Office of Naval Research
Arlington, VA 22217

15. SECURITY CLASS. (of this report) UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Educational test
Multiple-choice test
Mathematical models
Vocabulary measurement
k* index
Distractors

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This monograph reports research, related to the multiple-choice test item, which is conducted by psychometricians and educational technologists in Japan. Sato's number of hypothetical equivalent alternatives is introduced. The author proposes a new index, k*, which can be used, among other things, for invalidating three-parameter models for the multiple-choice item. Shiba's research on the measurement of vocabulary, which is based upon latent trait theory, includes an eventual tailored test on vocabulary, utilizing

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

404223 JV

20. Abstract (continued)

information obtained from distractors as well as correct answers. With this research in mind, the author has developed basic ideas about a new family of models for the multiple-choice item. These are based upon both the information given by distractors, and the correct answer and the noise resulting from random guessing.

RESEARCH ON THE MULTIPLE-CHOICE TEST ITEM IN JAPAN:

TOWARD THE VALIDATION OF MATHEMATICAL MODELS

FUMIKO SAMEJIMA

DEPARTMENT OF PSYCHOLOGY

UNIVERSITY OF TENNESSEE

KNOXVILLE, TN 37916

APRIL 1980

Prepared under the contract number N00014-77-C-0360
NR 150-402 with the
Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

RESEARCH ON THE MULTIPLE-CHOICE TEST ITEM IN JAPAN:

TOWARD THE VALIDATION OF MATHEMATICAL MODELS

ABSTRACT

This monograph reports research, related to the multiple-choice test item, which is conducted by psychometricians and educational technologists in Japan. Sato's number of hypothetical equivalent alternatives is introduced. The author proposes a new index, k*, which can be used, among other things, for invalidating three-parameter models for the multiple-choice item. Shiba's research on the measurement of vocabulary, which is based upon latent trait theory, includes an eventual tailored test on vocabulary, utilizing information obtained from distractors as well as correct answers. With this research in mind, the author has developed basic ideas about a new family of models for the multiple-choice item. These are based upon both the information given by distractors, and the correct answer and the noise resulting from random guessing.

PREFACE

In the summer of 1979, I spent a few weeks in Tokyo under the sponsorship of the Office of Naval Research (ONR). This monograph is based on conferences with researchers in Japan, in the areas of psychometrics, educational measurement, and educational technologies, and on research materials and technical literature collected during this trip. I thank Dr. Rudolph J. Marcus, Scientific Director, Miss Eunice Mohri, and other ONR/Tokyo staff members for providing me with office space and services, taking me to JICST, and helping me in many other ways.

I was invited to one of the bimonthly meetings of the Educational Technology Group of the Institute of Electronics and Communication Engineers in Japan, which was held at the Central Research Laboratories of Nippon Electric Co., Ltd., on 23 July, 1979, and had an opportunity to talk with the researchers who came to the meeting from many different districts of Japan. The author is thankful to Dr. Takahiro Sato, the representative of the Group, and other members for their kind cooperation in collecting research materials and literature.

It was also a pleasure to have several conferences with Dr. Sukeyori Shiba, Professor of Education at the University of Tokyo and an old friend of mine, during my stay in Tokyo, and to get to know a large scale research project on the measurement of vocabulary conducted by him and his students. The author is thankful to him and his students for making copies of their research materials and sending them to Knoxville, Tennessee, after I returned.

PREFACE (Continued)

Because of the shortage of time, the author could not see all the people she had wanted to; among them are Professor Takeuchi of the University of Tokyo and Dr. Akaike of the Institute of Mathematical Statistics, who happened to be out of town during her stay in Tokyo.

The stimulation of these conversations, and of the research materials and literature obtained in Tokyo, started new trains of thought in the author's mind. Some of these concern the multiple-choice item, which is the subject of this monograph. Others require yet more work and further communication with Japanese colleagues. In particular, the author feels it is worth trying to reanalyze the vocabulary test data collected by Shiba and others, using theory and methods which the author has developed and is going to develop.

The author is thankful to the Office of Naval Research for this opportunity of visiting Tokyo, and hopes that the present report will contribute to the development of mental test theory and science in general.

Fumiko Samejima

TABLE OF CONTENTS

I Introduction

There will not be any doubt in the mind of psychometricians that good mental test items are informative items, which make a great deal of contribution to the estimation of the examinee's ability, and, therefore, uncover the individual differences among the examinees accurately. In the history of mental test theory, the multiple-choice item arrived later than the free-response item, out of the necessity of administering group tests and of scoring their results speedily and objectively, in the sense that there is no need for our subjective judgment and evaluation in scoring. Today, an enormous number of multiple-choice tests are administered to youngsters, and their results have been used in many important decision-making situations, such as guidance, selection, classification, and so on. To construct good multiple-choice test items and to develop good mental test theory which deals with the multiple-choice item are, therefore, most important.

Since the multiple-choice item was introduced as a substitute for the free-response item, it has been treated by mental test theorists as something which is useful from the practical point of view, but not quite as good as the free-response item. The three-parameter logistic, or normal ogive, model, which is widely used by psychologists and educational psychologists for the multiple-choice item today, is nothing but a "blurred" image of the logistic, or normal ogive, model for the free-response item. In other words, there is nothing meaningful which is added to the original logistic, or normal ogive, model, but there are additional noises caused

by random guessing in the three-parameter logistic, or normal ogive model.

We must stop and think, however, if the three-parameter logistic, or normal ogive, model really fits psychological reality, and if the multiple-choice test item cannot be more than a "blurred" image of the free-response item. The author's answer to the first question is negative, to the second positive. It is clear in the author's mind that we need a better model than the three-parameter logistic, or normal ogive, model for the multiple-choice item, and that the multiple-choice item can provide us with a larger amount of information which results in a more accurate ability estimation, if we make use of the information given by its distractors, which the free-response item does not have.

It was interesting to discover that, while very few researchers in the United States have ever questioned the appropriateness of the three-parameter logistic, or normal ogive, model for the multiple-choice item, and have tried to validade it for their research data, the author's perception is shared by some Japanese researchers. Some of these are members of a nation-wide research group called the Educational Technology Group of the Institute of Electronics and Communication Engineers in Japan. Most of the members of the group are engineers in computer science, and some of them are educational psychologists. Tatsuoka has reported their names and research activities (Tatsuoka, 1979), which are represented by such topics as the S-P table (Student-Problem table),

the number of hypothetical, equivalent alternatives*, interpretive structural modeling based on graph theory, and so forth. Some of their papers, which the author has had the opportunity of reading, are listed in Appendix III. Their standpoint concerning the multiple-choice item is based on information theory (e.g., Goldman, 1953), considering that an item is a good one if its expected uncertainty in the selection of an alternative is high. As the measure of the quality of an item, the number of hypothetical, equivalent alternatives (Sato, 1977) is used, which will be introduced in Chapter 2. One impressive feature of the activities of this group of researchers is that they do not use computers mechanically, as many other researchers do, but they give teachers the feedback information about the test items constantly, and then they obtain the teachers' feedback based on the content analysis of the items in question, and so on. Another group is Shiba and his students of the School of Education, University of Tokyo. They have spent the past several years for developing vocabulary tests, which are aimed at measuring vocabulary of subjects of a wide range of age, collecting data, constructing an integrated vocabulary scale (Shiba, 1978), and then constructing a tailored test out of these vocabulary test items, using the information given by the distractors, as well as the correct answers, for branching examinees (Shiba, Noguchi and Haebara, 1978). The theory and method used for analyzing their data are basically the same as those adopted in the research in which the author was involved (Indow and Samejima, 1962, 1966).

---

*Tatusoka translated the original word as the effective (or equivalent) number of options, but the author uses this translation.

The outline of the work accomplished by Shiba and others will be given in Chapter 6.

With the research conducted by these people as incentives, the author has integrated her own ideas about mathematical models and the multiple-choice item. It resulted in proposing a method of validating, or invalidating, the three-parameter logistic, or normal ogive, model and the knowledge or random guessing principle, and eventually proposing a new family of models for the multiple-choice item, in which the information given by the distractors is fully utilized.

II Sato's Number of Hypothetical, Equivalent Alternatives

Let g $(=1,2,\ldots,n)$ be a multiple-choice test item. In the present paper, however, this symbol g is omitted, whenever it is clear that we deal with only one item. Let i $(=1,2,\ldots,m)$ be an alternative, or an option, of the multiple-choice item g, and p_i be the probability with which the examinee selects the alternative i. The entropy H is defined as the expectation of $-\log_2 p_i$ such that

$$(2.1) \qquad H = -\sum_{i=1}^{m} p_i \log_2 p_i ,$$

for the set of m alternatives of item g. It is obvious from (2.1) that the entropy H is non-negative, and, if one of the m alternatives is the sure event with unity as its probability, then $H = 0$. Sato's number of hypothetical, equivalent alternatives k, is defined by

$$(2.2) \qquad k = 2^H ,$$

and is used as an index of the effectiveness of the set of m alternatives for item g in the context of information theory. Since the entropy H indicates the expected uncertainty of the set of m events, or alternatives, the set of alternatives is more informative for a greater value of k.

When the probability p_i is replaced by the frequency ratio, P_i, we can write for the estimate of the entropy such that

$$(2.3) \qquad \hat{H} = -\sum_{i=1}^{m} P_i \log_2 P_i ,$$

and for the estimate of k we have

$$(2.4) \qquad \hat{k} = 2^{\hat{H}} .$$

We notice that we can obtain the number of hypothetical, equivalent alternatives k without using the entropy, for we have

$$(2.5) \qquad k = 2^{H} = 2^{-\sum_{i=1}^{m} p_i \log_2 p_i} = \prod_{i=1}^{m} p_i^{-p_i} = [\prod_{i=1}^{m} p_i^{p_i}]^{-1} .$$

The quantity in the brackets of the last expression of (2.5) is a kind of <u>weighted</u> <u>geometric</u> <u>mean</u> of p_i . Equation (2.5) also implies that we can use any base for $\log p_i$, instead of 2 . For convenience, hereafter we shall use e as the base of $\log p_i$, and use H^* instead of H such that

$$(2.6) \qquad H^* = -\sum_{i=1}^{m} p_i \log_e p_i \geqslant 0 \quad ,$$

which equals zero when one of the alternatives is the sure event, and

$$(2.7) \qquad k = e^{H^*} \geqslant 1 ,$$

and simply write $\log p_i$ instead of $\log_e p_i$.

To find out the value of p_i which maximizes H^* , and hence k , we define Q such that

$$(2.8) \qquad Q = -\sum_{i=1}^{m} p_i \log p_i + \lambda[\sum_{i=1}^{m} p_i - 1] ,$$

where λ is Lagrange's multiplier. Thus the partial derivative of Q with respect to p_i is given by

$$(2.9) \qquad \frac{\partial Q}{\partial p_i} = -[\log p_i + (1/p_i)p_i] + \lambda = -\log p_i + (\lambda - 1) .$$

Setting this derivative equal to zero, we obtain

$$(2.10)\qquad \log p_i = \lambda - 1 ,$$

which is a constant regardless of the value of i . Since we have

$$(2.11)\qquad \sum_{i=1}^{m} p_i = 1 ,$$

we obtain

$$(2.12)\qquad \hat{p}_i = 1/m .$$

Thus it is clear that H^* , and hence k , is maximal when all the m alternatives are equally probable, and we can write

$$(2.13)\qquad \max (H^*) = \log m$$

and

$$(2.14)\qquad \max (k) = m .$$

Since in the present situation the m events are alternatives, the values of H^* and k are affected by the difficulty level of item g . Let R be the correct answer to item g , which is given as one of its alternatives, and p_R be the probability with which the examinee selects the correct answer R . Figure 2-1 presents the relationship between the probability p_R and the number of hypothetical, equivalent alternatives k . In this figure, the area marked by slanted lines indicates the set of k's which are less than $\max (k|p_R)$ and greater than $\max[1/p_R, \min (k|p_R)]$, and are considered to be reasonable values of k by Sato and others.

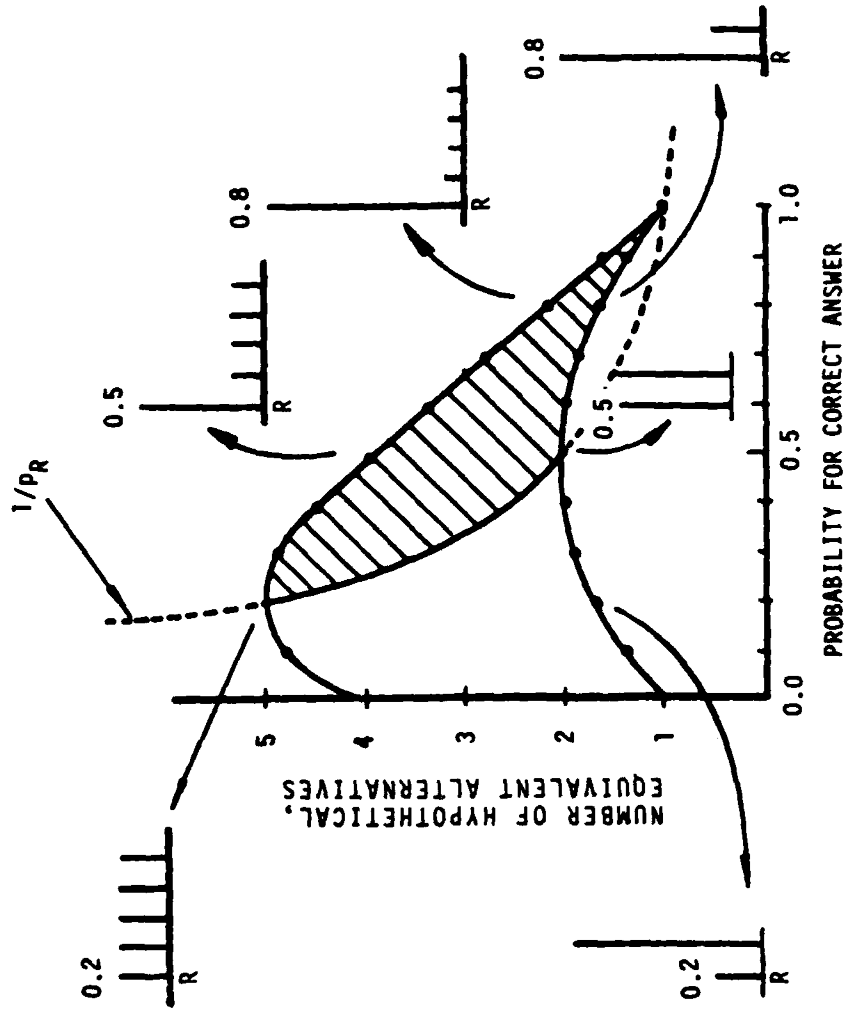


FIGURE 2-1

Relationship between the Probability with Which the Correct Answer R Is Selected and the Number of Hypothetical, Equivalent Alternatives, for Five-Choice Items.

In practice, Figure 2-1 is used by replacing the probability p_R by the proportion correct, P_R, and the number of hypothetical, equivalent alternatives, k, by its estimate $\hat{k}$. It is well-known that the frequency ratio is both the least squares solution and the maximum likelihood estimator of the corresponding probability. It is interesting to note that, in addition, it is the estimator which minimizes the chi-square statistic. Let us define Q such that

$$(2.15) \qquad Q = \sum_{i=1}^{m} [(NP_i - Np_i)^2/(Np_i)] + \lambda[\sum_{i=1}^{m} p_i - 1] ,$$

where N is the number of examinees and λ is Lagrange's multiplier. Then we have

$$(2.16) \qquad \frac{\partial Q}{\partial p_i} = N[(p_i^2 - P_i^2)/p_i^2] + \lambda = 0 ,$$

and

$$(2.17) \qquad \hat{p}_i = [1 + (\lambda/N)]^{-1/2} P_i .$$

Since

$$(2.18) \qquad 1 = \sum_{i=1}^{m} p_i = [1 + (\lambda/N)]^{-1/2} \sum_{i=1}^{m} P_i = [1 + (\lambda/N)]^{-1/2} ,$$

we obtain

$$(2.19) \qquad \lambda = 0 ,$$

and from this and (2.17) we can write

$$(2.20) \qquad \hat{p}_i = P_i .$$

The translation, "the number of hypothetical, equivalent alternatives," indicates the number of alternatives in the hypothetical situation where the entropy H is provided by the alternatives which are equivalent in the uncertainty of occurence. Although it is not the direct translation of the original word, it is used for k in the present paper, for it seems to the author to be the best describing word of the original.

III *Information Given by Distractors in the Multiple-Choice Item and Random Guessing*

Sato's number of hypothetical, equivalent alternatives has been used mainly by the members of the Technical Group of Educational Technologists in Japan (cf. Tatsuoka, 1979) for the purpose of analyzing the effectiveness of alternatives in relation with a relatively small group of examinees. The basic idea behind this index is that the expected uncertainty of the m events, or alternatives, be large, and, therefore, the number of hypothetical, equivalent alternatives be close to m. We notice that:

(1) this concept is strongly population-oriented, unlike those concepts in latent trait theory,

(2) it is assumed that each examinee tries to answer the item seriously, without depending upon random guessing,

and,

(3) relative to the population of examinees, the existence of too attractive a distractor is not desirable, since it tends to reduce the value of k.

Thus as long as this index is used for the analysis of test items which are given with careful guidance and supervision to samples of examinees from a well-defined population, and the findings of the analysis are not generalized across populations, it will serve its purpose.

If we generalize this concept and the resultant findings beyond these restrictions, however, we may be led to completely

false conclusions. To give an extreme example, suppose that none of our examinees took the test seriously, and selected one of the alternatives at random, for each item of the test. In such a case, regardless of the difficulty level of the item, the number of hypothetical, equivalent alternatives, k, will be very close to m for every item! In spite of this superficial success, we have obtained no information about the individual examinees' ability levels as the result of testing.

It is also noted that, if the examinee's behavior follows the knowledge or random guessing principle, i.e., he will answer correctly if he knows the answer, or guess randomly otherwise, the value of k tends to be large. In this case, too, our success of obtaining a large k is only superficial and meaningless.

In addition to the above facts, it is obvious that the value of the number of hypothetical, equivalent alternatives varies for different populations, i.e., the same item may have a value of k which is very close to m for one population of examinees, and may have a very low value for another population. This may be due to the difference in the mean ability levels of the two populations, or to the different forms of two ability distributions, or both. Thus while the index may be useful for a fixed population of examinees and if we discuss "how good an item is" in relation to that specific population, it cannot be considered as a parameter of the item per se. This limitation of the usefulness of k is of the same kind that is applicable for the reliability coefficient of the test, i.e., in spite of most psychologists' belief that

the reliability coefficient is one of the most important and solid properties of the test itself, it heavily depends upon the specific population of examinees for which the test is administered, and, therefore, is a dead concept since the population-free test information function is sufficient to serve the purpose (Samejima, 1977a).

As a whole, there is no single answer to the question: "Are items which have high values of the number of hypothetical, equivalent alternatives good items?" even if we control the testing situation with respect to the purpose of testing, such as guidance, selection, etc. This is true even if we restrict the populations of examinees, and it is mainly because of the noise induced by random guessing. That is to say, in a general situation of testing, it is hard for us to determine whether we have accomplished the work by obtaining a high value of k . In fact, the largest possible value of k may imply no accomplishment at all, as we have seen in one of the preceding paragraphs of the present chapter!

In spite of the above limitations, however, the introduction of the number of hypothetical, equivalent alternatives and its use by Sato and other researchers of the Technical Group of Educational Technologists should be well credited, for their vision is oriented toward the full use of the information given by all the alternatives of the multiple-choice item. It seems that they are quite successful in using the index in the small group situation, such as school classes where instructions are well conveyed and random guessing is extremely discouraged. This orientation is in quite a contrast to the attitude of many researchers who are accustomed

to the <u>blind use</u> of the three-parameter logistic model for the multiple-choice item, without ever stopping to think if the model can be validated for their data.

IV Three-Parameter Models in Latent Trait Theory and the Role of Item Distractors

Let θ be ability, or latent trait, that we intend to measure with our test. The three-parameter logistic model, or normal ogive model, is based upon the knowledge or random guessing principle, i.e., the examinee either knows the answer or guesses randomly. Let $\Psi_g(\theta)$ be the item characteristic function of item g, which is the conditional probability with which the examinee answers item g correctly, given θ, in the free-response situation. This is given by

$$(4.1) \qquad \Psi_g(\theta) = (2\pi)^{-1/2} \int_{-\infty}^{a_g(\theta-b_g)} e^{-u^2/2}\, du$$

in the normal ogive model, and

$$(4.2) \qquad \Psi_g(\theta) = [1 + \exp\{-Da_g(\theta-b_g)\}]^{-1}$$

in the logistic model, where a_g is the item discrimination parameter and b_g is the item difficulty parameter (Lord and Novick, 1968, Chapter 16), and D in (4.2) is the scaling factor which assumes 1.7 (Birnbaum, 1968) when the logistic model is used as a substitute for the normal ogive model.

The item characteristic function, $P_g(\theta)$, for the multiple-choice item in the three-parameter normal ogive, or logistic, model is defined by

$$(4.3) \qquad P_g(\theta) = \Psi_g(\theta) + [1-\Psi_g(\theta)]c_g = c_g + [1-c_g]\Psi_g(\theta) ,$$

where $\Psi_g(\theta)$ is given by (4.1) or (4.2) and c_g is a constant which

is called the guessing parameter, and equals $1/m_g$, or $1/m$.

It should be noted that, following these models, there is no information given by the alternatives other than the correct answer, for all the responses to the wrong answers are the result of random guessing. Should one of these models be valid for the item in question, the multiple-choice item would be nothing but a poor image of the binary, free-response item, which is contaminated by the noise caused by random guessing.

Let j be an individual examinee, and u_j be the binary item score for the multiple-choice item g . The conditional expectation and variance of the binary item score u , given θ , can be written as

$$(4.4) \qquad E(u|\theta) = P_g(\theta) = c + (1-c)\Psi_g(\theta) = (1/m)[1 + (m-1)\Psi_g(\theta)] ,$$

where c is the simplification of c_g , and

$$(4.5) \qquad \text{Var.}(u|\theta) = [(m-1)/m^2][1-\Psi_g(\theta)][1+(m-1)\Psi_g(\theta)] .$$

Let u_{ij} be the binary alternative score for the alternative i obtained by the individual j , for the multiple-choice item g . Thus we can write

$$(4.6) \qquad u_{Rj} = u_j .$$

The conditional expectation and variance of the binary alternative score u_i $(i \neq R)$, given θ , are given by

$$(4.7) \qquad E(u_i|\theta) = c[1-\Psi_g(\theta)] = (1/m)[1-\Psi_g(\theta)]$$

and

(4.8) $\quad \mathrm{Var.}(u_i|\theta) = (1/m^2)[1-\Psi_g(\theta)][(m-1)+\Psi_g(\theta)]$.

Let λ be either u or u_i , or any other discrete random variable, and $p(\lambda)$ and $p(\lambda|\theta)$ denote the marginal and conditional probability functions of λ , respectively. Then the relationships among the conditional and unconditional expectations and variances are given by

(4.9)
$$\begin{aligned} E(\lambda) &= \Sigma\ \lambda p(\lambda) = \Sigma\lambda \int_{-\infty}^{\infty} p(\lambda|\theta)f(\theta)d\theta = \int_{-\infty}^{\infty} \Sigma\ \lambda\ p(\lambda|\theta)f(\theta)d\theta \\ &= \int_{-\infty}^{\infty} E(\lambda|\theta)f(\theta)d\theta = E[E(\lambda|\theta)] \end{aligned}$$

and

(4.10)
$$\begin{aligned} \mathrm{Var.}(\lambda) &= \Sigma[\lambda-E(\lambda)]^2p(\lambda) = \Sigma[\lambda-E(\lambda)]^2 \int_{-\infty}^{\infty} p(\lambda|\theta)f(\theta)d\theta \\ &= \int_{-\infty}^{\infty} \Sigma[\lambda-E(\lambda|\theta)]^2p(\lambda|\theta)f(\theta)d\theta \\ &\qquad + \int_{-\infty}^{\infty} [E(\lambda|\theta)-E(\lambda)]^2\Sigma p(\lambda|\theta)f(\theta)d\theta \\ &= E[\mathrm{Var.}(\lambda|\theta)] + E[E(\lambda|\theta)-E(\lambda)]^2 . \end{aligned}$$

In particular, we can write

(4.11)
$$E(u) = E[E(u|\theta)] = \int_{-\infty}^{\infty} P_g(\theta)f(\theta)d\theta = p_R$$

and

(4.12)
$$\begin{aligned} \mathrm{Var.}(u) &= E[\mathrm{Var.}(u|\theta)] + E[E(u|\theta)-E(u)]^2 \\ &= \int_{-\infty}^{\infty} P_g(\theta)[1-P_g(\theta)]f(\theta)d\theta + \int_{-\infty}^{\infty} [P_g(\theta)-p_R]^2f(\theta)d\theta \\ &= p_R - p_R{}^2 = p_R(1-p_R) \end{aligned}$$

for the binary item score u , and, for the alternative score u_i ,

$$(4.13)\qquad E(u_i) = E[E(u_i|\theta)] = (1/m)\int_{-\infty}^{\infty}[1-\Psi_g(\theta)]f(\theta)d\theta$$
$$= [1/(m-1)]\int_{-\infty}^{\infty}[1-P_g(\theta)]f(\theta)d\theta = [1/(m-1)](1-p_R)$$
$$= p_i$$

and

$$(4.14)\qquad \mathrm{Var.}(u_i) = E[\mathrm{Var.}(u_i|\theta)] + E[E(u_i|\theta)-E(u_i)]^2$$
$$= (1/m^2)\int_{-\infty}^{\infty}[1-\Psi_g(\theta)][(m-1)+\Psi_g(\theta)]f(\theta)d\theta$$
$$+ (1/m^2)\int_{-\infty}^{\infty}[\{1-\Psi_g(\theta)\}-mp_i]^2 f(\theta)d\theta$$
$$= (1/m)\int_{-\infty}^{\infty}[1-\Psi_g(\theta)]f(\theta)d\theta$$
$$- 2p_i(1/m)\int_{-\infty}^{\infty}[1-\Psi_g(\theta)]f(\theta)d\theta + p_i^2$$
$$= p_i(1-p_i)\ .$$

We notice that $E(u)$ given in (4.11) is the item difficulty parameter in classical test theory, which depends upon the specific population of examinees as well as the test item.

It should be noted that both the expectation and the variance of u_i for $i \neq R$, which are given by (4.13) and (4.14), respectively, are equal for all the wrong answers, and are determined, solely, by p_R and the number of the alternatives, m. This is the logical consequence of the fact that the responses to those wrong answers are completely the result of random guessing, and provide us with no information about the examinees' ability levels.

We must remember, however, that most of the conscientious test constructors try to avoid the contamination of the quality of items, by finding <u>incorrect, but plausible</u>, answers and including them as <u>distractors</u> in the set of alternatives. This indicates

that the responses to these alternatives are not the result of random guessing, and may contain useful information about the examinee's ability level. The adoption of one of the three-parameter models for such multiple-choice items is not justifiable, since in so doing the researchers distort psychological reality and will produce nothing but meaningless artifacts as the result of their research.

It is strange to the author that many researchers have ignored the contradiction which was described in the preceding paragraphs, and have applied the three-parameter models to their data for years, which, obviously, are based on the tests containing many distractors. As far as they continue repeating this mistake, their conscientiousness as researchers has to be questioned.

V Index k^* for Invalidating Three-Parameter Models

It has been pointed out in Chapter 3 that Sato's number of hypothetical, equivalent alternatives takes on a high value, if every examinee in the group has selected one of the m alternatives at random. This fact implies that, although the index was introduced for quite an opposite purpose, it may also be useful in detecting the examinee's random guessing behavior in the multiple-choice item.

To materialize the above, we need the following consideration. When the examinee follows the knowledge or random guessing principle and the item characteristic function assumes the three-parameter logistic, or normal ogive, model, the index k is solely affected by the probability with which the examinee knows the answer, as is obvious from Figure 2-1 and (4.3) and (4.11). This fact provides some inconvenience, however, for the probability of knowing the answer heavily depends upon the specific population of examinees, in addition to the item characteristic function of the item in the free-response situation. It will be more convenient, therefore, if we can modify Sato's index k in such a way that it is unaffected by the ability distribution of a specific population of examinees, and can be considered as a pure property of the item. With this aim in mind, we shall introduce a new index in this chapter.

Let $\bar{A}$ be the event that the examinee does not know the answer to item g, and consider the probability space which consists of such a subpopulation of examinees. The conditional probability, $p(i|\bar{A})$, with which the examinee selects the alternative

i of item g in this conditional probability space is given by

$$(5.1) \qquad p(i|\bar{A}) \begin{cases} = p_i [\sum_{i \neq R} p_i + p_R^*]^{-1} & i \neq R \\ = p_R^* [\sum_{i \neq R} p_i + p_R^*]^{-1} , & i = R \end{cases}$$

where p_R^* denotes the probability with which the examinee guesses correctly for item g . The new index, k^* , is defined in terms of these conditional probabilities, in such a way that

$$(5.2) \qquad k^* = \exp[-\sum_{i=1}^{m} p(i|\bar{A}) \cdot \log p(i|\bar{A})] = [\prod_{i=1}^{m} p(i|\bar{A})^{p(i|\bar{A})}]^{-1} .$$

It is obvious that $p(i|\bar{A})$ for $i \neq R$ is proportional to p_i , for every examinee in the population who has selected one of the wrong answers does not know the answer, and, consequently, he is also in the subpopulation $\bar{A}$. On the other hand, examinees who have selected the correct answer R are not necessarily in the subpopulation $\bar{A}$, so we can write

$$(5.3) \qquad p_R^* \leqslant p_R .$$

Note that, if the examinee's behavior follows the knowledge or random guessing principle and the item characteristic function of the multiple-choice item g is of one of the three-parameter models, p_R^* equals p_i for $i \neq R$, and, as the result, all the m $p(i|\bar{A})$'s are equal and $k^* = m$.

In practice, we need to use some estimates for $p(i|\bar{A})$'s , to obtain the estimate of k^* . Since we have the frequency ratio, P_i , for the estimate of p_i for $i \neq R$, all we need to do is to

find out an appropriate estimate of p_R^*. Let P_R^* denote such an estimate of p_R^*, and P_i^* be such that

$$(5.4) \qquad P_i^* \begin{cases} = P_i & i \neq R \\ = P_R^* & i = R . \end{cases}$$

Then we can write for the estimate of $p(i|\bar{A})$ such that

$$(5.5) \qquad \hat{p}(i|\bar{A}) = P_i^* [\sum_{i=1}^{m} P_i^*]^{-1} .$$

We are to take the strategy of finding P_R^* which makes k^* maximal. Define $\hat{H}^*$ such that

$$(5.6) \qquad \hat{H}^* = \log \hat{k}^* = - \sum_{i=1}^{m} \hat{p}(i|\bar{A}) \cdot \log \hat{p}(i|\bar{A})$$

$$= -[\sum_{s=1}^{m} P_s^*]^{-1} [\sum_{i=1}^{m} P_i^* \cdot \log P_i^* - (\sum_{i=1}^{m} P_i^*) \cdot \log \{\sum_{s=1}^{m} P_s^*\}] .$$

Then the partial derivative of $\hat{H}^*$ with respect to P_R^* can be written as

$$(5.7) \qquad \frac{\partial \hat{H}^*}{\partial P_R^*} = [\sum_{s=1}^{m} P_s^*]^{-2} [\sum_{i=1}^{m} P_i^* \cdot \log P_i^* - (\sum_{s=1}^{m} P_s^*) \cdot \log P_R^*] ,$$

and, setting this equal to zero, we obtain

$$(5.8) \qquad \log P_R^* = [\sum_{s \neq R} P_s]^{-1} \sum_{i \neq R} P_i \cdot \log P_i$$

and then

$$(5.9) \qquad P_R^* = \prod_{i \neq R} P_i^{P_i [\sum_{s \neq R} P_s]^{-1}}$$

Thus we can use (5.9) in (5.4), and, therefore, obtain $\hat{p}(i|\bar{A})$

through (5.5). The estimate of the new index, k^* , is given by

$$(5.10) \qquad \hat{k}^* = \exp[-\sum_{i=1}^{m} \hat{p}(i|\bar{A})\cdot\log \hat{p}(i|\bar{A})] = [\prod_{i=1}^{m} \hat{p}(i|\bar{A})^{\hat{p}(i|\bar{A})}]^{-1} .$$

A necessary, though not sufficient, condition for one of the three-parameter models to be valid is that $\hat{k}^*$ should be equal to m within sampling fluctuations, <u>regardless of the population of examinees</u> from which our sample happened to be selected. If this is not the case, we must say that the three-parameter model does not fit our item, i.e., the invalidation of the model.

Although the invalidation of the three-parameter logistic, or normal ogive, model is easy, its validation is more difficult. We recall that Sato's number of hypothetical, equivalent alternatives is used as a measure of the desirability of the item for a specific population of examinees. If all the distractors are equally probable for a specific population, then the index k^* will also equal m , in spite of the fact that the two cases are completely different in nature. This problem can be solved by administering the same test to a different group of examinees, which has a different ability distribution from that of the first group. If the large value of k^* is due to the knowledge or random guessing principle, then it will also be large for the second group of examinees because of its population-free nature. On the other hand, if the large value of k^* is resulted from the optimal quality of the item for the first group of examinees, then it will not be as large as that for the second group, unless the operating characteristics of all the distractors are identical.

It should be emphasized that k^* takes on a large value even if the knowledge or random guessing principle does not work behind the examinee's behavior, but the item is "suitable" for the group of examinees to which the test has been administered, in the same sense that a high value of Sato's number of hypothetical, equivalent alternatives is meant to indicate. This fact means that, when we need to use only one set of data for validating, or invalidating, the knowledge or random guessing principle and the three-parameter logistic, or normal ogive, model, we must use, at least, one more necessary condition for the principle to be valid. One such necessary condition is that the sample means of ability θ , or of its estimate, of the subgroups of examinees who have selected the wrong answers should be equal, within the range of sampling fluctuations. Thus, if <u>either</u> the value of k^* is substantially less than m , <u>or</u> the sample means of ability θ of such subgroups of examinees are not close to each other, then we shall be able to say that the knowledge or random guessing principle and the three-parameter model are invalidated. On the other hand, if both of the necessary conditions are satisfied with our data, we can say there is no reason to reject the principle and the model.

For the purpose of illustration, a set of simulated data was calibrated, using the Monte Carlo method. In this set of data, five hypothetical multiple-choice test items were assumed, each having five alternatives, A, B, C, D and E, with A always as the correct answer. Each item is assumed to follow the three-parameter normal ogive model, which is given by (4.1) and (4.3), with the parameter values shown in Table 5-1. A group of five hundred

TABLE 5-1

Item Discrimination Parameter a_g and Item Difficulty Parameter b_g of Each of the Five Hypothetical, Binary Items Following the Three-Parameter Normal Ogive Model, with $c_g = 0.2$.

Item	a_g	b_g
1	1.00	0.00
2	1.50	0.00
3	2.00	0.00
4	2.50	0.00
5	3.50	0.00

hypothetical examinees was assumed, whose ability levels are placed at one hundred equally spaced points on the ability continuum, which start with -2.475 and end with 2.475, in such a way that subjects 1 through 5 are placed at $\theta = -2.475$, subjects 6 through 10 are at $\theta = -2.425$, and so on. For each of the five hypothetical multiple-choice items, the response of each of the five hundred hypothetical examinees was calibrated according to the specified item characteristic function and the knowledge or random guessing principle. These calibrated responses are presented as Table A-1 in Appendix I.

Table 5-2 presents the frequency ratio, P_i , of each of the five alternatives, for each of the five hypothetical multiple-choice items. We can see that sampling fluctuations are fairly large for item 4, and to a less degree for item 2, since the corresponding probability, p_i , is 0.6 for the alternative A and 0.1 for each of the alternatives B, C, D and E. In the same table, also presented are the values of P^*_R , which were obtained through (5.9). Using these values in (5.6), (5.9) and (5.10), the estimates of the entropy H^* and the index k^* were obtained, and are presented in Table 5-3. Since the maximal possible value of $\hat{H}^*$ is approximately 1.60944 (=log m) and that of $\hat{k}^*$ is 5 (=m), we can say that these results are sufficiently close to their respective maximal values, i.e., an exemplification of the satisfaction of one of the necessary conditions for validating the three-parameter normal ogive model and the knowledge or random guessing principle by our simulated data. The fact that these results are less

TABLE 5-2

Frequency Ratio of the Subject, P_i , Who Selected Each of the Five Alternatives, and the Modified Frequency Ratio P_R^* for the Correct Answer A, for Each of the Five Hypothetical Items.

Item \ Alternative		A	B	C	D	E
1	P_i	.608	.086	.106	.100	.100
	P_R^*	.098				
2	P_i	.618	.102	.080	.106	.094
	P_R^*	.096				
3	P_i	.600	.094	.106	.108	.092
	P_R^*	.100				
4	P_i	.606	.104	.078	.130	.082
	P_R^*	.101				
5	P_i	.598	.092	.100	.104	.106
	P_R^*	.101				

TABLE 5-3

Entropy, $\hat{H}^*$, and the Number of Hypothetical, Equivalent Alternatives, $\hat{k}^*$, for Each of the Five Hypothetical Items Following the Three-Parameter Normal Ogive Model.

Item	$\hat{H}^*$	$\hat{k}^*$
1	1.60714	4.98853
2	1.60501	4.97789
3	1.60744	4.99000
4	1.59224	4.91475
5	1.60829	4.99424

satisfactory for item 4 and the same is true, to a lesser degree, for item 2 must be due to the sampling fluctuations, which were observed in Table 5-2.

As another necessary condition for validating the three-parameter normal ogive model and the knowledge or random guessing principle, the mean of θ for each of the five subgroups of examinees, who selected different alternatives, was computed, for each of the five multiple-choice items. Table 5-4 presents the result of these means of θ. In the same table, also presented is the expectation of θ for each of the five subgroups, using the uniform ability distribution for the interval, [-2.5, 2.5], for each item, following the *three-parameter normal* ogive model and the knowledge or random guessing principle. Since all the responses to one of the four wrong answers of each item are nothing but the result of random guessing, these alternatives are equivalent, and have the same mean value of θ. We can see that, for each item, the mean of θ for the correct answer and that of each incorrect answer are substantially different, and they are close enough to the respective theoretical means.

In practice, there is no way to observe the examinee's θ itself. We can use its maximum likelihood estimate, $\hat{\theta}$, however, and use it as the substitute in the above process, for example. We must obtain a similar result as above, to validate the three-parameter models and the knowledge or random guessing principle.

We notice that a similar result as the one in our example

TABLE 5-4

Sample Mean of θ for the Subgroup of Hypothetical Examinees Who Selected Each of the Five Alternatives, and Its Corresponding Theoretical Mean, for Each of the Five Multiple-Choice Items.

Alternative / Item	A (Correct) $E(\theta)$	A (Correct) $\bar{\theta}$	B $\bar{\theta}$	C $\bar{\theta}$	D $\bar{\theta}$	E $\bar{\theta}$	(Incorrect) $E(\theta)$
1	0.703	0.619	-0.912	-1.017	-0.994	-0.905	-1.054
2	0.774	0.752	-1.341	-1.084	-1.249	-1.161	-1.161
3	0.800	0.811	-1.165	-1.233	-1.224	-1.237	-1.200
4	0.812	0.809	-1.230	-1.119	-1.253	-1.369	-1.218
5	0.822	0.809	-1.061	-1.193	-1.260	-1.282	-1.234

can be obtained, if, incidentally, all the distractors require "on the average" approximately the same level of ability for the examinee to be attracted to them, for our group of examinees. This fact indicates that it is desirable to add more necessary conditions to examine, such as the approximate equality of the second moment of θ , or $\hat{\theta}$, that of the third moment, etc., for the subgroups of examinees who have selected the wrong answers. Since these subgroups of examinees are "equivalent" in ability distribution if the knowledge or random guessing principle and the three-parameter model are valid, these higher moments should be equal within sampling fluctuations, which it is highly unlikely that all the subgroups of examinees who have been attracted to separate distractors are equivalent in ability distribution. We must avoid, however, using moments of too high degrees, for their sampling fluctuations tend to be enormously great.

VI <u>Shiba's Research on the Measurement of Vocabulary</u>

In this chapter, we shall introduce a research on the measurement of vocabulary, which was conducted by Shiba and others. The author found it interesting, especially in the following aspects.

(1) The vocabulary tests they used are very well constructed, choosing each alternative carefully.

(2) Subjects were selected from many different age groups.

(3) Unlike many researchers in the United States, they have tried to make a full use of the distractors.

The battery of tests used for the construction of the vocabulary scale consists of eleven tests, A1, A2, A3, A4, A5, A6, J1, J2, S1, S2 and U . Each test contains thirty to fifty-eight multiple-choice items, each having a set of five alternatives. These tests differ in difficulty, and each of them is designed for a different group of ages, ranging from six years of age to the ages of college students. There are subsets of items included in two tests, which are adjacent to each other in difficulty. For example, items 37 through 56 of Test J1 are also items 1 through 20 of Test J2. The number of examinees used for the vocabulary scale construction varies between 412 sixth graders of elementary schools for Test A5 and 924 second graders of senior high schools for Test S1. (cf. Shiba, 1978.)

The model adopted for the item characteristic function of each vocabulary item is the logistic model, such that

$$(6.1) \qquad P_g(\theta) = [1 + \exp\{-Da_g(\theta-b_g)\}]^{-1} ,$$

where a_g and b_g are the item discrimination and difficulty parameters, respectively, and $D = 1.7$. Note that Shiba did not use the three-parameter logistic model, which is characterized by (4.2) and (4.3). This is based on his belief that three-parameter models are not applicable for well-developed multiple-choice items, which he has formed through his many experiences in test construction and research.

Each of the eleven tests was administered to a group of subjects who belong to a single school year, except for college students. Hereafter, for convenience, we shall use EL for elementary schools, JH for junior high schools, SH for senior high schools, and CS for colleges, and add the school year after each symbol. For instance, by SH2 we mean a group of subjects who are in the second year of senior high schools. The correspondence of the subject groups and the tests administered is summarized as follows:

A1 for EL1 (650), A2 for EL2 (650), A3 for EL3 (546),
A4 for EL4 (617), A5 for EL5 (599), A6 for EL6 (412),
J1 for JH1 (614), J2 for JH2 (758), S1 for SH1 (924),
S2 for SH2 (759) and U for CS (740) ,

where the numbers in parentheses indicate respective numbers of examinees. Note that JH3 and SH3 are not included in the data which are the basis of the vocabulary scale construction.

The main steps for analyzing these data are the following.

[A] For each of the eleven groups of examinees, the ability distribution is assumed to be the standard normal distribution.

[B] Assuming the normal ogive model, such that

(6.2) $$P_g(\theta) = (2\pi)^{-1/2}\int_{-\infty}^{a_g(\theta-b_g)} e^{-u^2/2}\,du ,$$

where a_g and b_g are the item discrimination and difficulty parameters, respectively, and the local independence of the item variables (Lord and Novick, 1968, Chapter 16), and also that the regression of each item variable on ability θ is linear, the tetrachoric correlation coefficient is computed for each and every pair of items.

[C] The principal factor solution of factor analysis is applied for the correlation matrix thus obtained, using the largest absolute value of the correlation coefficient in each row, or column, as the communality. This step is also the process of validating the uni-dimensionality of ability θ . Figure 6-1 illustrates the resulting set of eigenvalues for Test J1 which was administered to 614 first year junior high school students. It turned out that the first eigenvalue is much larger than all the other eigenvalues, and thus the uni-dimensionality was confirmed. Hereafter, this first principal factor is treated as θ .

[D] From the result of factor analysis, the item parameters are obtained. Let ρ_g be the factor loading (e.g., Lawley and Maxwell, 1971) of the first principal factor, or θ , for item g . The item discrimination parameter, a_g , is obtained by

(6.3) $$a_g = \rho_g(1-\rho_g)^{-1/2} .$$

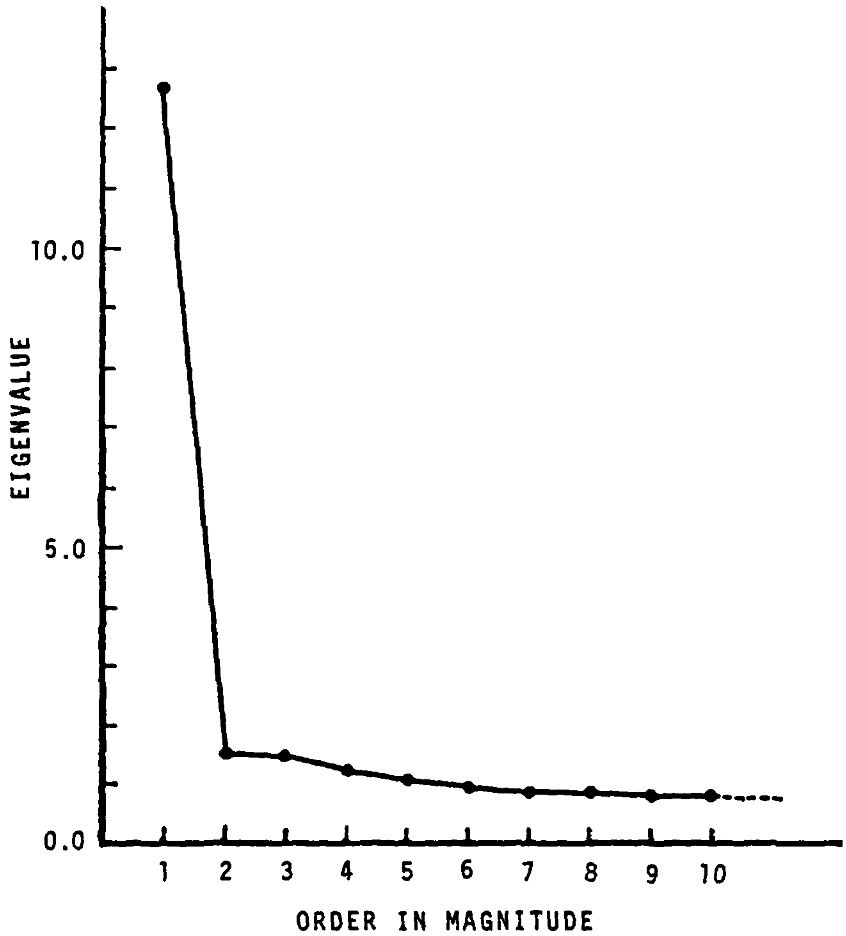


FIGURE 6-1

Eigenvalues of the Correlation Matrix of the Fifty-Five Items of Test J1, Ordered with Respect to Their Magnitudes.

Let $\Phi(u)$ denote the standard normal distribution function, such that

$$(6.4) \qquad \Phi(u) = (2\pi)^{-1/2} \int_{-\infty}^{u} e^{-t^2/2}\, dt \, .$$

The item difficulty parameter, b_g , is given by

$$(6.5) \qquad b_g = \Phi^{-1}(1-p_{gr})\, \rho_g^{-1} \, ,$$

where p_{gR} is the probability with which the examinee answers item g correctly. In practice, this is replaced by the frequency ratio, P_{gR} , to provide us with the estimate of b_g .

[E] The eleven ability scales thus constructed are considered to be on the same continuum, and they are integrated into a single scale. This equating is made through the ten subsets of items, each of which is shared by two adjacent tests. Let a_g and b_g be the item parameters estimated from the result of the first test, and a_g^* and b_g^* be those from the result of the second test. Denoting the two ability scales by θ and θ^* , respectively, we can write

$$(6.6) \qquad a_g(\theta-b_g) = a_g^*(\theta^*-b_g^*) \, ,$$

since the item characteristic functions, which follow the normal ogive model, of the same item g on the two ability scales must assume the same value for the corresponding values of θ and θ^* . Thus the functional relationship between

θ and θ^* is given by

(6.7) $$\theta^* = (a_g/a_g^*)\theta + [b_g^* - (a_g/a_g^*)b_g] ,$$

which is linear, and the two coefficients are obtained from these four parameters. In practice, we obtain as many sets of coefficients as the number of common items, and we need to use some type of "average" of these coefficients for the scale transformation. Figure 6-2 presents the ability distributions of the eleven subject groups after such transformations were made and the mean and the standard deviation of the distribution of J1 are taken as the origin and the unit for the new, integrated ability dimension.

[F] The item characteristic function of each item on the new, integrated scale θ is approximated by the logistic function, which is given by (6.1).

[G] The maximum likelihood estimate, $\hat{\theta}_j$, of each examinee's ability is obtained through the equation

(6.8) $$\sum_{g=1}^{n} a_g P_g(\hat{\theta}_j) = \sum_{g=1}^{n} a_g u_{gj}$$

(cf. Birnbaum, 1968), where u_{gj} is the binary item score of individual j for item g .

[H] The test information function of each test is obtained by

(6.9) $$I(\theta) = \sum_{g=1}^{n} I_g(\theta) ,$$

where $I_g(\theta)$ is the item information function of item g such

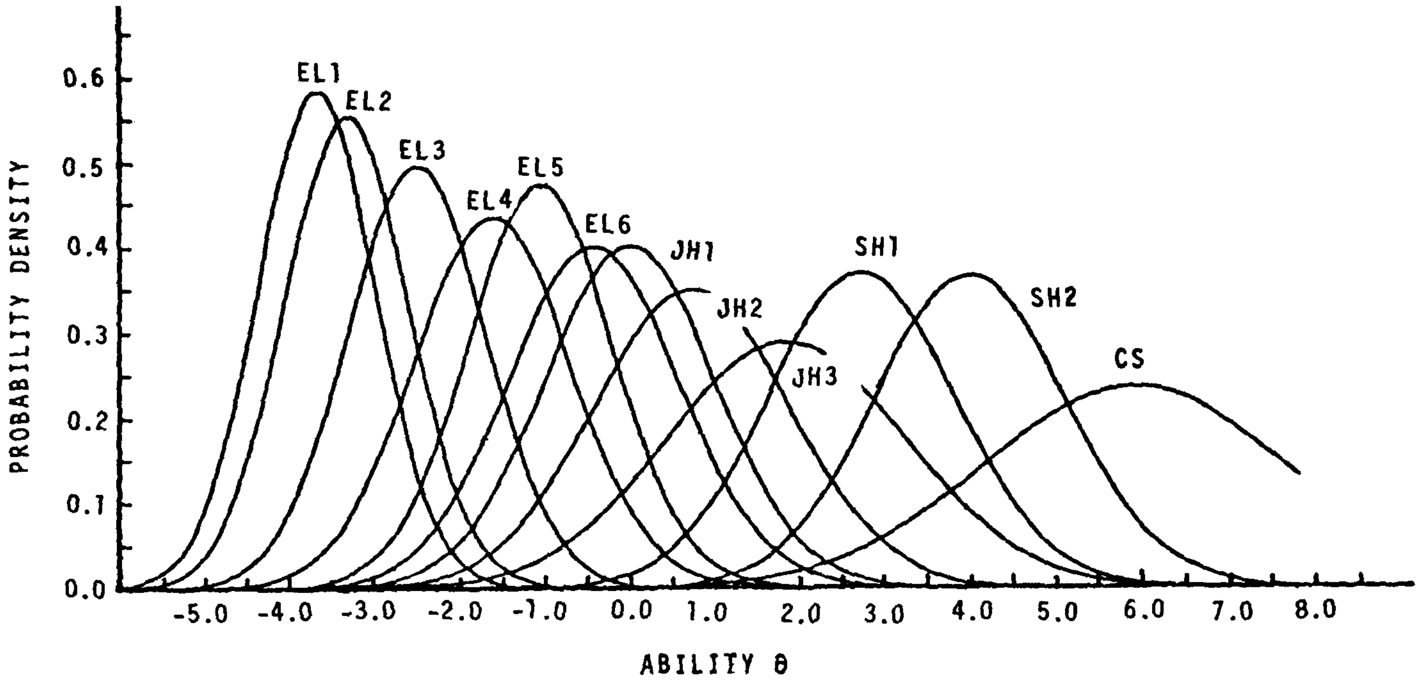


FIGURE 6-2

Estimated Density Functions of the Twelve Groups of Examinees, Which Are Assumed to Be Normal. The Ability Scale Is Defined in Such a Way that the Density Function of the First Grade Group of Junior High School (JH1) Is n(0,1) .

that

$$(6.10) \qquad I_g(\theta) = [P_g'(\theta)]^2 [P_g(\theta)\{1-P_g(\theta)\}]^{-1} .$$

Figure 6-3 presents the test information functions thus obtained for the eleven tests.

[I] The theoretical frequency distribution of test score T for each test and examinee group can be written as

$$(6.11) \qquad N \sum_{V \in T} \sum_{u_g \in V} P_g(\theta)^{u_g} [1-P_g(\theta)]^{1-u_g} ,$$

where V is a response pattern or a vector of n item scores, and T is the test score given by

$$(6.12) \qquad T = \sum_{g=1}^{n} u_g .$$

This is used for the validation of the model and assumptions adopted in the process of analysis. Figure 6-4 illustrates the goodness of fit of this theoretical frequency distribution of test score to the actual frequency distribution, for Test J1.

[J] The sample mean of the maximum likelihood estimate $\hat{\theta}$ of the subgroup of examinees, who selected each of the five alternatives is calculated, for each item of each test.

[K] A tailored test of the vocabulary is constructed by selecting an appropriate subset of items from these eleven tests, in such a way that an individual is directed to a next item which is chosen on the basis of the sample mean of $\hat{\theta}$ of the alternative

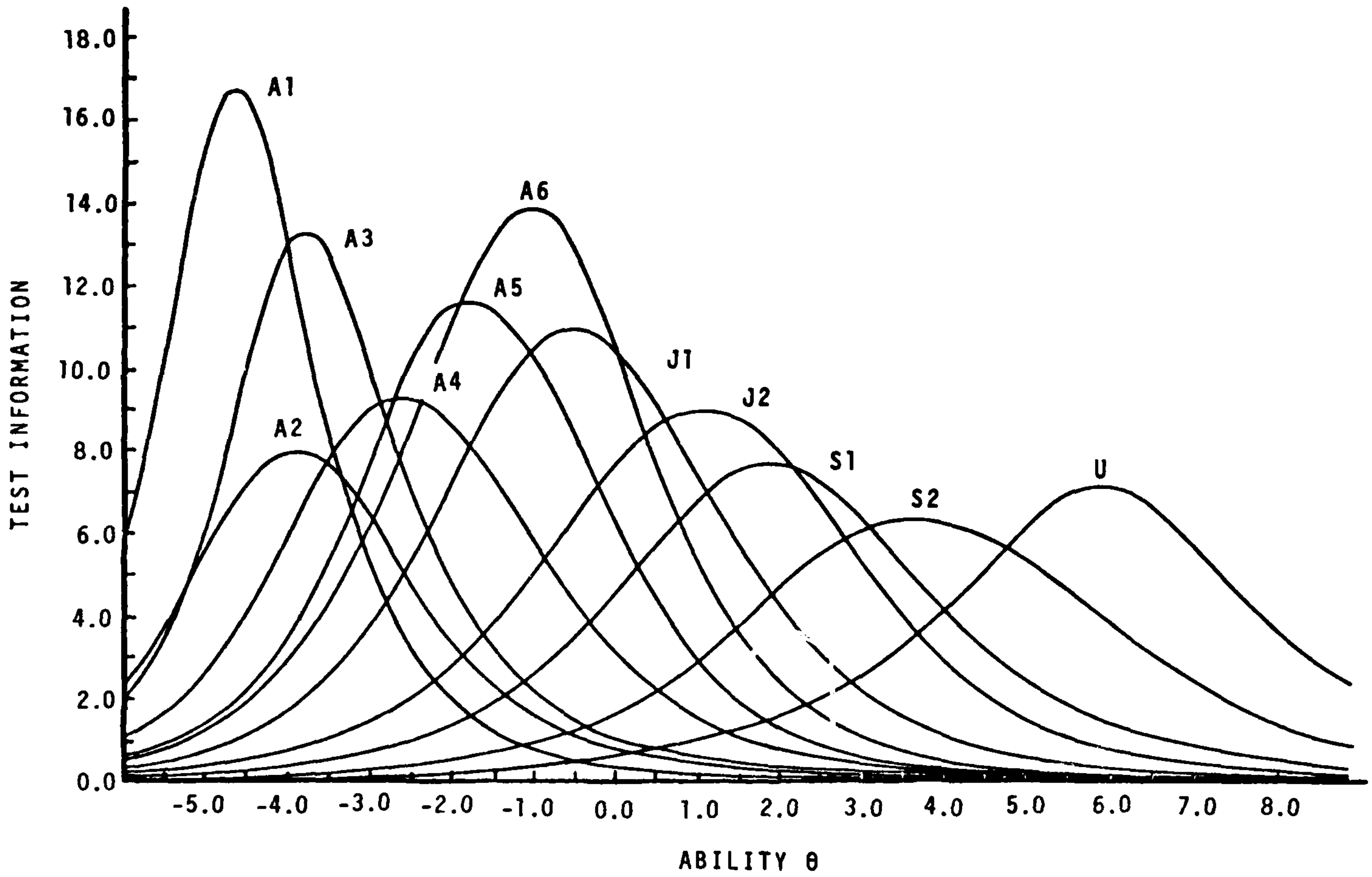


FIGURE 6-3

Test Information Functions of the Eleven Tests on Vocabulary.

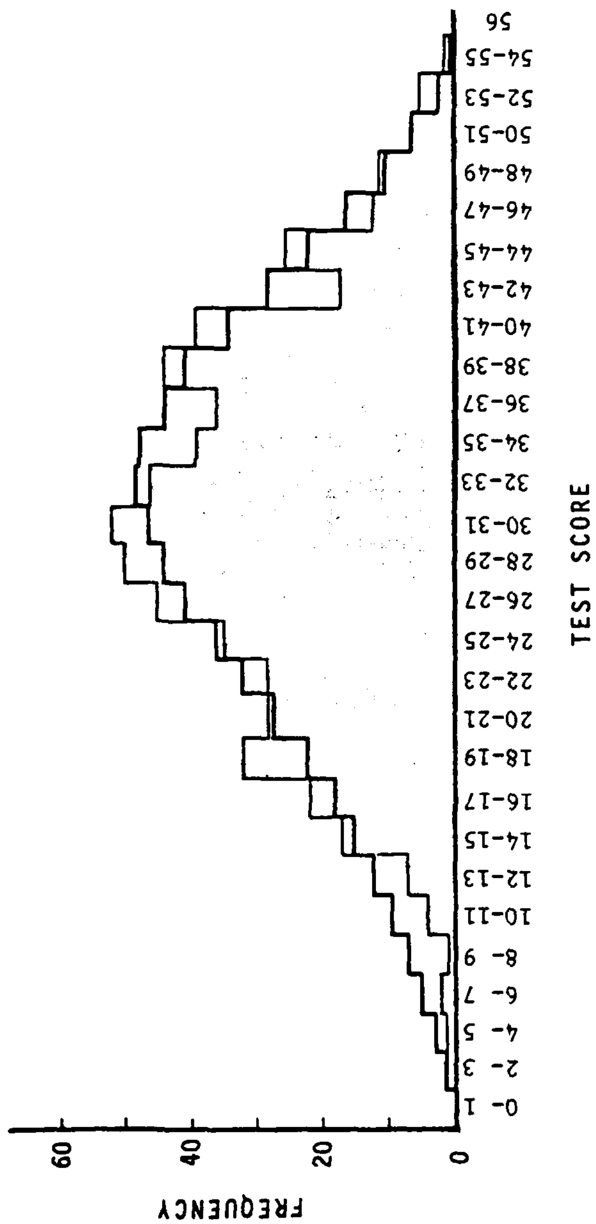


FIGURE 6-4

Theoretical and Observed (shaded) Frequency Distributions of the Test Score of Test J1.

he has selected for the present item.

We have seen in the preceding paragraphs a brief sketch of Shiba and others' work. It is unfortunate that the author cannot convey the fine quality of the tests themselves to the reader, for they are vocabulary tests and their translation from Japanese into English would certainly destroy the nature of the tests. We can see that the research has been conducted very conscientiously, however, including several processes of validation, and has eventually produced a widely applicable vocabulary scale and a tailored test. In the latter result, although there is some room for improvement, the use of distractors for "branching" subjects should be taken as a stimulation to the researchers who are engaged in this area, for it has seldom been seriously investigated by other researchers.

The research conducted by Shiba and others includes more interesting data than were used in the vocabulary scale construction. Table 6-1 presents a part of them, in which the frequency distribution of the alternative selection and the mean of the maximum likelihood estimate of ability for each alternative are shown for nineteen items included in both Tests J1 and J2, and administered to four different subject groups, JH1, JH2(a), JH2(b) and JH3. In the same table, also presented is the discrepancy between the mean of $\hat{\theta}$ for the correct answer and the lowest mean $\hat{\theta}$ for one of the four wrong answers, under the heading, "largest discrepancy." The correct answers are always identified as the ones which have the highest means of $\hat{\theta}$, except for the one for item 8 administered to JH2(b), which is the second highest

TABLE 6-1

Mean of the Maximum Likelihood Estimates of Ability, $\hat{\theta}$, for Each of the Five Subgroups of Subjects Selecting Different Alternatives, for Each of the 19 Vocabulary Test Items, Together with the Actual Frequency Distributions (FRQ). The Difference between the Mean $\hat{\theta}$ of the Correct Subgroups and the Lowest Mean $\hat{\theta}$ Is Also Presented As Largest Discrepancy for Each Item. Test J1, Junior High School Grade 1

Item	Indices	Alternative 1	2	3	4	5	Total	Largest Discrepancy
37	Mean $\hat{\theta}$	0.401	-0.476	-0.482	-0.750	-0.148	572	1.151
	FRQ	287	50	59	59	117		
38	Mean $\hat{\theta}$							
	FRQ							
39	Mean $\hat{\theta}$	-0.192	-0.091	-0.270	-0.243	0.400	562	0.670
	FRQ	91	115	118	51	187		
40	Mean $\hat{\theta}$	0.071	-0.416	-0.336	0.310	-0.479	573	0.789
	FRQ	60	141	90	273	9		
41	Mean $\hat{\theta}$	-0.557	-1.007	-0.445	-0.456	0.254	573	1.261
	FRQ	53	20	23	85	392		
42	Mean $\hat{\theta}$	0.339	-0.570	0.036	-0.439	-0.387	570	0.909
	FRQ	247	21	121	84	97		
43	Mean $\hat{\theta}$	-0.512	0.376	-0.572	-0.245	-0.393	572	0.948
	FRQ	26	308	98	67	73		
44	Mean $\hat{\theta}$	-0.293	-0.547	-0.595	0.271	-0.318	569	0.866
	FRQ	119	67	14	333	36		
45	Mean $\hat{\theta}$	-0.638	-0.412	-0.636	0.395	-0.593	568	1.033
	FRQ	51	25	123	346	23		
46	Mean $\hat{\theta}$	0.444	-0.741	-0.325	-0.428	-0.534	568	1.185
	FRQ	296	46	44	164	18		
47	Mean $\hat{\theta}$	-0.261	0.270	-0.078	-0.426	-0.101	569	0.696
	FRQ	69	224	158	53	65		
48	Mean $\hat{\theta}$	-0.129	-0.024	-1.013	-0.467	0.412	564	1.425
	FRQ	81	100	58	67	258		
49	Mean $\hat{\theta}$	-0.339	-0.390	-0.284	-0.464	0.309	573	0.773
	FRQ	115	31	42	70	315		
50	Mean $\hat{\theta}$	0.349	-0.256	-1.015	-0.317	-0.385	571	1.364
	FRQ	308	46	35	86	96		
51	Mean $\hat{\theta}$	-0.137	-0.640	-0.077	-0.136	0.429	560	1.069
	FRQ	89	82	75	113	201		
52	Mean $\hat{\theta}$	-0.219	0.291	-0.110	-0.608	-0.095	565	0.899
	FRQ	116	235	80	34	100		
53	Mean $\hat{\theta}$	-0.071	-0.030	-0.453	0.527	-0.241	572	0.980
	FRQ	163	51	34	143	181		
54	Mean $\hat{\theta}$	0.132	-0.060	-0.084	-0.037	-0.283	561	0.415
	FRQ	182	111	100	142	26		
55	Mean $\hat{\theta}$	0.114	-0.278	-0.172	-0.533	0.690	571	1.223
	FRQ	27	72	317	29	126		
56	Mean $\hat{\theta}$	-0.460	-0.113	-0.412	0.742	0.015	572	1.202
	FRQ	104	101	115	141	111		

TABLE 6-1 (Continued): Test J1, Junior High School Grade 2

Item	Indices	Alternative 1	2	3	4	5	Total	Largest Discrepancy
37	Mean $\hat{\theta}$	0.886	-0.215	-0.249	-0.312	0.028	455	1.198
	FRQ	269	39	39	37	71		
38	Mean $\hat{\theta}$							
	FRQ							
39	Mean $\hat{\theta}$	0.384	0.186	0.083	-0.068	1.015	450	1.083
	FRQ	55	97	82	50	166		
40	Mean $\hat{\theta}$	0.521	-0.133	0.109	0.802	-0.286	458	1.088
	FRQ	61	95	45	243	14		
41	Mean $\hat{\theta}$	-0.553	-0.440	-0.173	-0.019	0.665	461	1.218
	FRQ	27	13	19	47	355		
42	Mean $\hat{\theta}$	0.810	-0.426	0.348	-0.089	-0.201	457	1.236
	FRQ	257	14	68	67	51		
43	Mean $\hat{\theta}$	-0.162	0.791	-0.578	0.142	-0.321	458	1.369
	FRQ	10	312	53	46	37		
44	Mean $\hat{\theta}$	0.298	-0.145	-0.228	0.664	0.237	456	0.892
	FRQ	65	54	15	291	31		
45	Mean $\hat{\theta}$	-0.124	0.139	-0.290	0.823	-0.469	459	1.292
	FRQ	30	23	79	299	28		
46	Mean $\hat{\theta}$	0.849	-0.751	-0.263	-0.260	-0.072	459	1.600
	FRQ	308	25	29	90	7		
47	Mean $\hat{\theta}$	-0.136	0.764	-0.119	-0.194	-0.001	459	0.958
	FRQ	43	302	54	30	30		
48	Mean $\hat{\theta}$	0.483	0.262	-0.889	-0.086	0.871	455	1.760
	FRQ	56	85	38	45	231		
49	Mean $\hat{\theta}$	0.050	-0.351	0.183	-0.419	0.756	460	1.175
	FRQ	96	16	19	35	294		
50	Mean $\hat{\theta}$	0.798	0.153	-0.634	0.151	-0.099	455	1.432
	FRQ	269	19	20	84	63		
51	Mean $\hat{\theta}$	0.118	-0.260	0.312	0.150	0.909	448	1.169
	FRQ	76	47	55	68	202		
52	Mean $\hat{\theta}$	0.195	0.778	0.035	0.206	0.177	449	0.743
	FRQ	60	239	71	21	58		
53	Mean $\hat{\theta}$	0.376	0.193	-0.013	0.918	0.040	459	0.931
	FRQ	94	34	26	180	125		
54	Mean $\hat{\theta}$	0.817	0.256	0.282	0.221	0.051	451	0.766
	FRQ	177	75	82	108	9		
55	Mean $\hat{\theta}$	-0.043	-0.042	-0.052	-0.455	1.157	458	1.612
	FRQ	20	45	174	18	201		
56	Mean $\hat{\theta}$	0.256	0.236	-0.289	1.354	0.247	455	1.643
	FRQ	70	100	80	128	77		

TABLE 6-1 (Continued): Test J2, Junior High School Grade 2

Item	Indices	Alternative 1	2	3	4	5	Total	Largest Discrepancy
1	Mean $\hat{\theta}$ FRQ	-0.247 145	-0.901 11	-1.148 19	-1.354 11	-0.744 35	221	1.107
2	Mean $\hat{\theta}$ FRQ							
3	Mean $\hat{\theta}$ FRQ	-0.667 28	-0.660 45	-0.639 42	-0.834 16	-0.224 87	218	0.610
4	Mean $\hat{\theta}$ FRQ	-0.403 51	-0.963 30	-1.036 23	-0.289 115	-0.948 2	221	0.747
5	Mean $\hat{\theta}$ FRQ	-1.126 14	-1.573 2	-1.070 10	-1.091 18	-0.334 177	221	1.239
6	Mean $\hat{\theta}$ FRQ	-0.239 125	-0.948 6	-0.607 32	-0.891 32	-0.978 25	220	0.739
7	Mean $\hat{\theta}$ FRQ	-2.089 1	-0.269 153	-1.365 24	-0.671 30	-0.946 13	221	1.820
8	Mean $\hat{\theta}$ FRQ	-0.761 37	-1.205 12	-0.589 6	-0.376 156	-0.362 10	221	0.829
9	Mean $\hat{\theta}$ FRQ	-1.259 10	-0.746 9	-1.098 21	-0.312 172	-1.428 8	220	1.116
10	Mean $\hat{\theta}$ FRQ	-0.194 141	-1.057 11	-0.850 18	-1.096 47	-0.648 4	221	0.902
11	Mean $\hat{\theta}$ FRQ	-1.035 22	-0.253 143	-0.801 26	-1.059 7	-0.924 22	220	0.806
12	Mean $\hat{\theta}$ FRQ	-0.681 23	-0.883 23	-1.551 10	-1.113 18	-0.251 147	221	1.300
13	Mean $\hat{\theta}$ FRQ	-0.597 50	-1.016 6	-0.777 21	-1.277 13	-0.302 131	221	0.975
14	Mean $\hat{\theta}$ FRQ	-0.227 134	-0.860 13	-1.523 9	-0.646 34	-1.023 30	220	1.296
15	Mean $\hat{\theta}$ FRQ	-0.766 34	-1.045 18	-0.845 26	-0.974 36	-0.061 107	221	0.984
16	Mean $\hat{\theta}$ FRQ	-0.764 36	-0.093 87	-0.571 54	-1.369 11	-0.784 29	217	1.276
17	Mean $\hat{\theta}$ FRQ	-0.704 52	-0.373 21	-0.858 5	-0.128 85	-0.842 58	221	0.730
18	Mean $\hat{\theta}$ FRQ	-0.189 109	-0.745 33	-0.731 31	-0.929 39	-0.291 7	219	0.740
19	Mean $\hat{\theta}$ FRQ	-1.012 5	-0.808 38	-0.875 88	-1.139 7	0.148 83	221	1.287
20	Mean $\hat{\theta}$ FRQ	-0.923 46	-0.805 38	-0.948 46	0.304 67	-0.507 24	221	1.252

TABLE 6-1 (Continued): Test J2, Junior High School Grade 3

Item	Indices	Alternative 1	2	3	4	5	Total	Largest Discrepancy
1	Mean $\hat{\theta}$	0.161	-0.838	-0.787	-1.099	-0.374	573	1.260
	FRQ	436	30	25	19	63		
2	Mean $\hat{\theta}$							
	FRQ							
3	Mean $\hat{\theta}$	-0.312	-0.287	-0.373	-0.486	0.351	567	0.837
	FRQ	54	93	97	63	260		
4	Mean $\hat{\theta}$	-0.025	-0.848	-0.252	0.181	-0.709	572	1.029
	FRQ	83	77	38	362	12		
5	Mean $\hat{\theta}$	-0.763	-0.766	-0.864	-0.611	0.107	574	0.971
	FRQ	30	7	19	43	475		
6	Mean $\hat{\theta}$	0.221	-0.722	-0.267	-0.675	-0.801	568	1.022
	FRQ	371	7	96	49	45		
7	Mean $\hat{\theta}$	-0.597	0.175	-1.125	-0.339	-0.870	570	1.300
	FRQ	10	441	45	50	24		
8	Mean $\hat{\theta}$	-0.438	-0.966	-0.448	0.100	-0.272	571	1.066
	FRQ	55	31	14	457	14		
9	Mean $\hat{\theta}$	-1.089	-0.368	-0.828	0.252	-0.780	570	1.341
	FRQ	32	47	67	407	17		
10	Mean $\hat{\theta}$	0.117	-1.019	-0.229	-1.035	0.022	571	1.152
	FRQ	473	15	28	51	4		
11	Mean $\hat{\theta}$	-0.555	0.264	-0.750	-0.666	-0.619	572	1.014
	FRQ	36	389	69	35	43		
12	Mean $\hat{\theta}$	-0.478	-0.511	-1.394	-0.754	0.203	572	1.597
	FRQ	33	87	10	35	407		
13	Mean $\hat{\theta}$	-0.595	-0.888	-0.366	-1.342	0.216	573	1.558
	FRQ	107	16	29	14	407		
14	Mean $\hat{\theta}$	0.241	-0.367	-1.527	-0.382	-0.824	571	1.768
	FRQ	387	22	12	84	66		
15	Mean $\hat{\theta}$	-0.610	-0.853	-0.582	-0.638	0.441	561	1.294
	FRQ	67	27	79	69	319		
16	Mean $\hat{\theta}$	-0.629	0.264	-0.499	-0.344	-0.555	569	0.893
	FRQ	58	364	75	14	58		
17	Mean $\hat{\theta}$	-0.277	0.166	-0.469	0.351	-0.546	572	0.897
	FRQ	109	42	30	259	132		
18	Mean $\hat{\theta}$	0.383	-0.380	-0.418	-0.548	-0.579	565	0.962
	FRQ	294	65	80	115	11		
19	Mean $\hat{\theta}$	-0.943	-0.582	-0.651	-0.789	0.439	572	1.382
	FRQ	15	87	136	9	325		
20	Mean $\hat{\theta}$	-0.524	-0.484	-0.770	0.692	-0.363	574	1.462
	FRQ	78	74	93	235	94		

in mean $\hat{\theta}$.

VII Use of Index k^* When Distractors Are in Full Work

It is obvious in Table 6-1 of the preceding chapter that for these vocabulary items the knowledge or random guessing principle does not work behind the examinee's behavior, for the mean values of $\hat{\theta}$ for the wrong answers are substantially different from one another for most of the items. In cases like this, index k^*, which was introduced in Chapter 5 as a modification of Sato's number of hypothetical, equivalent alternatives and used as an index for invalidating three-parameter models, can be used as a measure of desirability of the item for the group of examinees in question, just as Sato's index is meant to be used for. An additional merit of index k^* when it is used for this purpose will be that it can be used directly, without depending upon the relationship with the probability for the correct answer, p_R, which is illustrated by Figure 2-1.

Table 7-1 presents the estimated entropy $\hat{H}^*$ obtained by (5.6), for each of the nineteen items and each of the four groups of examinees, JH1, JH2(a), JH2(b) and JH3. The values of index $\hat{k}^*$, which correspond to these $\hat{H}^*$'s in Table 7-1, were obtained by (5.10) and are shown in Table 7-2.

We can see in these tables that thirteen out of the total of nineteen items have higher values of $\hat{H}^*$, and hence of $\hat{k}^*$, for JH2(a) than for JH2(b). Since the subjects in these two groups are of the same school year, i.e., the second year of junior high school, this tendency may be related with the fact that for JH2(a) these nineteen items were given at the end of the test and for

TABLE 7-1

Entropy of Each of the Nineteen Vocabulary Items Based on Each of the Four Subgroups, i.e., Junior High School, Grades 1, 2, 2 and 3. For the First Two Subgroups of Subjects Test J1 Was Used and for the Other Two Subgroups Test J2 Was Used.

Subgroup / Item	JH1	JH2(a)	JH2(b)	JH3
37 (1)	1.55907	1.57572	1.51218	1.52080
39 (3)	1.57359	1.57997	1.55547	1.53566
40 (4)	1.41987	1.48141	1.39913	1.46917
41 (5)	1.47880	1.52098	1.46885	1.48496
42 (6)	1.50740	1.51576	1.50880	1.42679
43 (7)	1.54070	1.51224	1.39256	1.49871
44 (8)	1.43049	1.51333	1.41791	1.47934
45 (9)	1.42195	1.49895	1.54177	1.52485
46(10)	1.37234	1.36152	1.36544	1.39912
47(11)	1.52673	1.58391	1.54137	1.57599
48(12)	1.59254	1.57072	1.57317	1.43130
49(13)	1.51299	1.40124	1.40700	1.32933
50(14)	1.54630	1.46214	1.50665	1.43095
51(15)	1.59962	1.59600	1.58320	1.55950
52(16)	1.54651	1.54903	1.51294	1.51407
53(17)	1.45244	1.46629	1.41821	1.48312
54(18)	1.51192	1.45933	1.51052	1.46054
55(19)	1.23002	1.27989	1.25075	1.30371
56(20)	1.60838	1.60223	1.58595	1.60504

TABLE 7-2

Number of Hypothetical, Equivalent Alternatives of Each of the Nineteen Vocabulary Items Based on Each of the Four Subgroups, i.e., Junior High School, Grades 1, 2, 2 and 3. For the First Two Subgroups of Subjects Test J1 Was Used and for the Other Two Subgroups Test J2 Was Used.

Subgroup / Item	JH1	JH2(a)	JH2(b)	JH3
37 (1)	4.75440	4.83420	4.53660	4.57590
39 (3)	4.82391	4.85480	4.73730	4.88252
40 (4)	4.13659	4.39917	4.05166	4.34565
41 (5)	4.38768	4.57672	4.34425	4.41479
42 (6)	4.51496	4.55290	4.52130	4.16531
43 (7)	4.66784	4.53688	4.02513	4.47592
44 (8)	4.18076	4.54183	4.12850	4.39004
45 (9)	4.14519	4.47701	4.67284	4.59447
46(10)	3.94459	3.90212	3.91744	4.05162
47(11)	4.60310	4.87397	4.67098	4.83551
48(12)	4.91623	4.81011	4.82191	4.18412
49(13)	4.54029	4.06023	4.08370	3.77850
50(14)	4.69408	4.31519	4.51161	4.18267
51(15)	4.95113	4.93326	4.87053	4.75646
52(16)	4.69506	4.70690	4.54008	4.54521
53(17)	4.27352	4.33314	4.12972	4.40669
54(18)	4.53542	4.30307	4.52908	4.30829
55(19)	3.42128	3.59625	3.49295	3.68295
56(20)	4.99472	4.96410	4.88392	4.97805

JH2(b) they were given at the beginning of the test. We can also observe that, for some items, there exists a mild tendency that the value of $\hat{k}^*$ becomes greater as the school year increases, and, for some others, this tendency is reversed. Items 39(3), 40(4), 44(8), 45(9), 47(11), 53(17) and 55(19) belong to the first category, and items 37(1), 48(12), 49(13), 50(14), 51(15), 52(16) and 54(18) are members of the second category. In spite of these mild tendencies, however, the values of index $\hat{k}^*$ are large, ranging, approximately, from 3.42 to 4.99 , for all the examinee groups, the result which indicates a high desirability of this subset of test items for these groups of examinees.

We can observe a tendency that, regardless of the groups of examinees, some items have higher values of $\hat{k}^*$ than others, and some other items have lower values of $\hat{k}^*$ than others. Items 56(20), 51(15) and 39(3) exemplify the first category, and items 55(19) and 46(10) are members of the second category.

The mean and the standard deviation of the nineteen values of $\hat{k}^*$ for each of the four examinee groups were computed, and are presented in Table 7-3. We can see that all the mean values are between 4.39 and 4.51, and all the standard deviations are between 0.34 and 0.40, i.e., very close to one another, respectively.

As an additional information, the product-moment correlation coefficient of $\hat{k}^*$'s , which are shown in Table 7-2, was computed for each pair of examinee groups, and the result is presented in Table 7-4. We can see that these values are fairly large and positive, as we can expect from Table 7-2.

TABLE 7-3

Mean and Standard Deviation (s.d.) of the Index k* for the Nineteen Vocabulary Items, for Each of the Four Examinee Groups.

Examinee Group	Mean	s.d.
JH1	4.4832	0.3944
JH2(a)	4.5038	0.3659
JH2(b)	4.3931	0.3759
JH3	4.3976	0.3465

TABLE 7-4

Product-Moment Correlation Coefficient of the Index k* for Each Pair of the Four Examinee Groups.

	JH1	JH2(a)	JH2(b)	JH3
JH1	1.00000	0.82705	0.82711	0.60447
JH2(a)	0.82705	1.00000	0.85120	0.85770
JH2(b)	0.82711	0.85120	1.00000	0.71444
JH3	0.60447	0.85770	0.71444	1.00000

The result of the principal factor analysis of the correlation matrix, Table 7-4, with the largest correlation coefficient of each row or column as the first estimate of the communality and using three iterative reestimations of the communalities, provides us with the eigenvalues, 3.237, 0.266, 0.044 and -0.011 . Since the correlation matrix, with communalities as the principal diagonal elements, is positive semi-definite, the negative eigenvalue is due to the error, resulting, mainly, from the inaccuracy of the estimation of the communalities. The final communality estimates are approximately 0.863, 0.999, 0.862 and 0.833, respectively. We can say from this result that a strong, dominating general factor exists behind the four sets of $\hat{k}^*$'s , since the first eigenvalue, 3.237, is by far the largest, and the other eigenvalues are close to zero. The first factor loadings for the four examinee groups, which are the correlation coefficients between this general factor and the separate sets of $\hat{k}^*$'s , respectively, turned out to be 0.868, 0.983, 0.905 and 0.836 .

These facts indicate that the four examinee groups are fairly similar to one another with respect to the configuration of the values of $\hat{k}^*$ as far as these nineteen vocabulary test items are concerned.

VIII Proposal of a New Family of Models for the Multiple-Choice Item

Throughout the history of mental measurement, the multiple-choice item has been treated as a "poor image of the free-response item," and very little accomplishment has been made in pursuing its theoretical advantage, rather than its handicap. Most researchers in these days mechanically adopt the three-parameter logistic model for their research which is based on the multiple-choice item, without even trying to validate the model. As long as they continue doing this, we shall never be able to expect any progress in this area of science, in spite of the fact that more and more research materials and published papers are accumulated year by year.

It has been one of the author's purposes of pursuing the method of estimating the operating characteristics without assuming any mathematical model a priori (Samejima, 1977b, 1977c, 1978a, 1978b, 1978c, 1978d, 1978e, 1978f) to approach the operating characteristics of distractors, which are completely neglected by the users of three-parameter models. While this approach is undoubtedly more scientific than any others, it will be desirable to consider new types of models, which reflect psychological reality behind the examinee's behavior in the multiple-choice situation far better than three-parameter models and the knowledge or random guessing principle.

The research on the vocabulary measurement made by Shiba and others should be credited for the fact that they did not

accept the fashionable three-parameter logistic model blindly as many other researchers do, and, moreover, they try to make full use of the information given by the distractors to the extent that they used it for branching examinees in tailored testing. As far as we treat the multiple-choice item as a binary item, it will be a poor substitute for the free-response item, which is contaminated by noise or guessing. If we make use of the information given by the distractors, however, the multiple-choice item can be more informative than the free-response item, and will no longer be a poor image of the free-response item.

The family of models that will be proposed in this chapter is related with the graded response model (Samejima, 1969, 1972), in which an item is scored into more than two response categories. Let x_g be the graded item score, which assumes integers, 0 through m_g, and $P_{x_g}(\theta)$ be its operating characteristic. The graded response level can be classified into the homogeneous and the heterogeneous cases (Samejima, 1972), and we can name the normal ogive model (Samejima, 1972, 1973) and the logistic model (Samejima, 1972) as models in the homogeneous case, and Bock's multi-nomial response model (Bock, 1972, Samejima, 1972) as an example in the heterogeneous case. In these models, the operating characteristic of the item response category is defined, respectively, as follows.

$$(8.1) \qquad P_{x_g}(\theta) = (2\pi)^{-1/2} \int_{a_g(\theta-b_{x_g+1})}^{a_g(\theta-b_{x_g})} e^{-u^2/2}\, du \ .$$

(8.2) $$P_{x_g}(\theta) = [1+\exp\{-Da_g(\theta-b_{x_g})\}]^{-1} - [1+\exp\{-Da_g(\theta-b_{x_g+1})\}]^{-1}.$$

(8.3) $$P_{x_g}(\theta) = \exp\{\alpha_{x_g}\theta+\beta_{x_g}\}[\sum_{s=o}^{m_g}\exp\{\alpha_s\theta+\beta_s\}]^{-1}.$$

In both the normal ogive and the logistic models, i.e., in (8.1) and (8.2), the item parameter a_g is a positive number, and the item response parameter b_{x_g} satisfies the relationship such that

(8.4) $$-\infty = b_0 < b_1 < b_2 < \ldots < b_{m_g} < b_{m_g+1} = \infty \quad .$$

In the latter, D is a positive number which assumes 1.7 when the logistic model is used as a substitute for the normal ogive model. In Bock's multi-nomial model, one of the item response parameters, α_{x_g} satisfies the inequality,

(8.5) $$\alpha_0 \lesseqgtr \alpha_1 \lesseqgtr \alpha_2 \lesseqgtr \cdots\cdots \lesseqgtr \alpha_{m_g} \quad .$$

Suppose that the multiple-choice item g is constructed in such a way that all the main, plausible answers are covered by the alternatives, in addition to the correct answer. Suppose, further, that no guessing is involved in the examinee's behavior in answering item g . Then the examinee will either be attracted to one of the alternatives, or will have no idea at all as to its answer. Arrange all the distractors in the order of their plausibility, and give the numbers 1 through (m_g-1) in the ascending order. The number assigned to the correct answer is m_g , or m for simplicity, and the one assigned to the "no idea at all" category

is 0 . In such a situation, the operating characteristic of the graded response category can be used as the operating characteristic of the alternative, treating "no answer" as the additional alternative, to which the item score is 0 .

In practice, however, because of the pressure of testing, it is rather unlikely that the examinee will leave the item unanswered even when he has "no idea at all." For this reason, now we shall assume that the examinee guesses randomly when he is not attracted by the plausibility of any alternative. Thus we shall deal with the m alternatives as the graded response categories, 1 through m , and we can write for the operating characteristic of the alternative

$$(8.6) \qquad P_{x_g}(\theta) = \Psi_{x_g}(\theta) + (1/m_g)[1-\sum_{s=1}^{m_g}\Psi_s(\theta)] \ , \quad x_g=1,2...,m_g \ ,$$

where $\Psi_{x_g}(\theta)$ is the operating characteristic of the alternative which is numbered x_g , when no guessing is involved. Thus we can use one of the $P_{x_g}(\theta)$'s defined by (8.1), (8.2) and (8.3), or a similar operating characteristic of the graded response category with a sound rationale behind it, depending upon the nature of the item and the set of alternatives.

For the purpose of illustration, we shall use the normal ogive model for $\Psi_{x_g}(\theta)$, with $a_g = 1.5$ and b_{x_g}'s are -2.0, -1.0, 0.0, 1.0 and 2.0 for $x_g = 1,2,3,4,5$, respectively. Figure 8-1 presents the operating characteristics of the (m_g+1) alternatives, obtained by (8.1), when no guessing is involved and "no answer" is treated as the additional alternative, or category 0 .

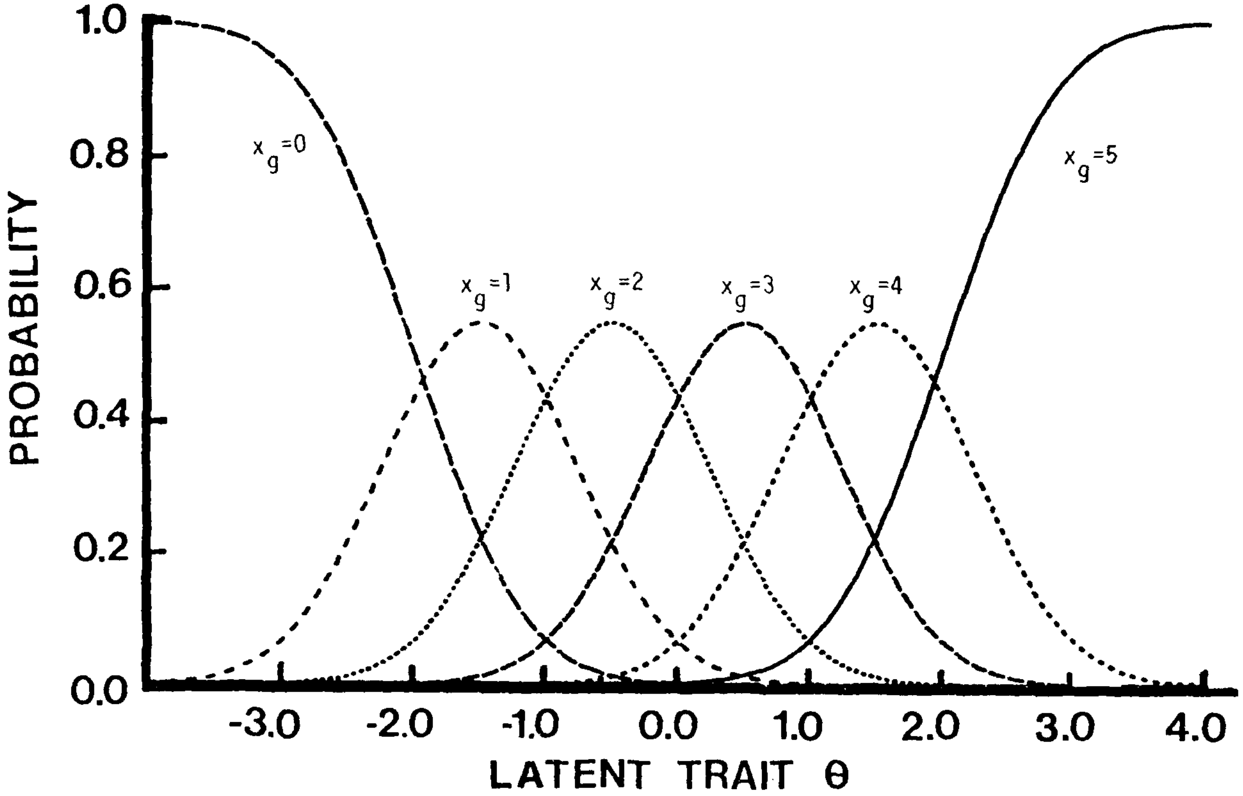


FIGURE 8-1

Operating Characteristics of Six Item Response Categories Following the Normal Ogive Model, with $a_g = 1.5$, $b_1 = -2.0$, $b_2 = -1.0$, $b_3 = 0.0$, $b_4 = 1.0$ and $b_5 = 2.0$.

In this example, the operating characteristics of the four distractors are unimodal, with -1.5, -0.5, 0.5 and 1.5 as the modal points, respectively. Figure 8-2 presents the operating characteristics of the five alternatives when guessing is involved, which are given by (8.6) with $\Psi_{x_g}(\theta)$ replaced by $P_{x_g}(\theta)$ given in (8.1). We can see that, unlike the operating characteristics when no guessing is involved, these curves have the common asymptote, $1/5$, when θ approaches negative infinity. To compare the two operating characteristics of each alternative more clearly, Figure 8-3 presents the two curves for each alternative in one graph, with the dotted line for the one without guessing, and the solid line for the one with guessing.

The family of models presented by (8.6) seems reasonable, in the sense that it considers both the information given by the distractors and the noise caused by random guessing. Its behavior will be investigated further, and will be discussed in a separate paper.

It is interesting to note that the use of the normal ogive model and its logistic approximation in the research on vocabulary measurement conducted by Shiba and others can be justified by the new family of models. As we can see in the fifth graph of Figure 8-3, when the parameter b_1 is as distant from b_{m_g} as in this example, the operating characteristic of the correct answer is practically the same as the item characteristic function of the normal ogive model on the dichotomous response level, except for the additional "tail" on the lower levels of ability. If

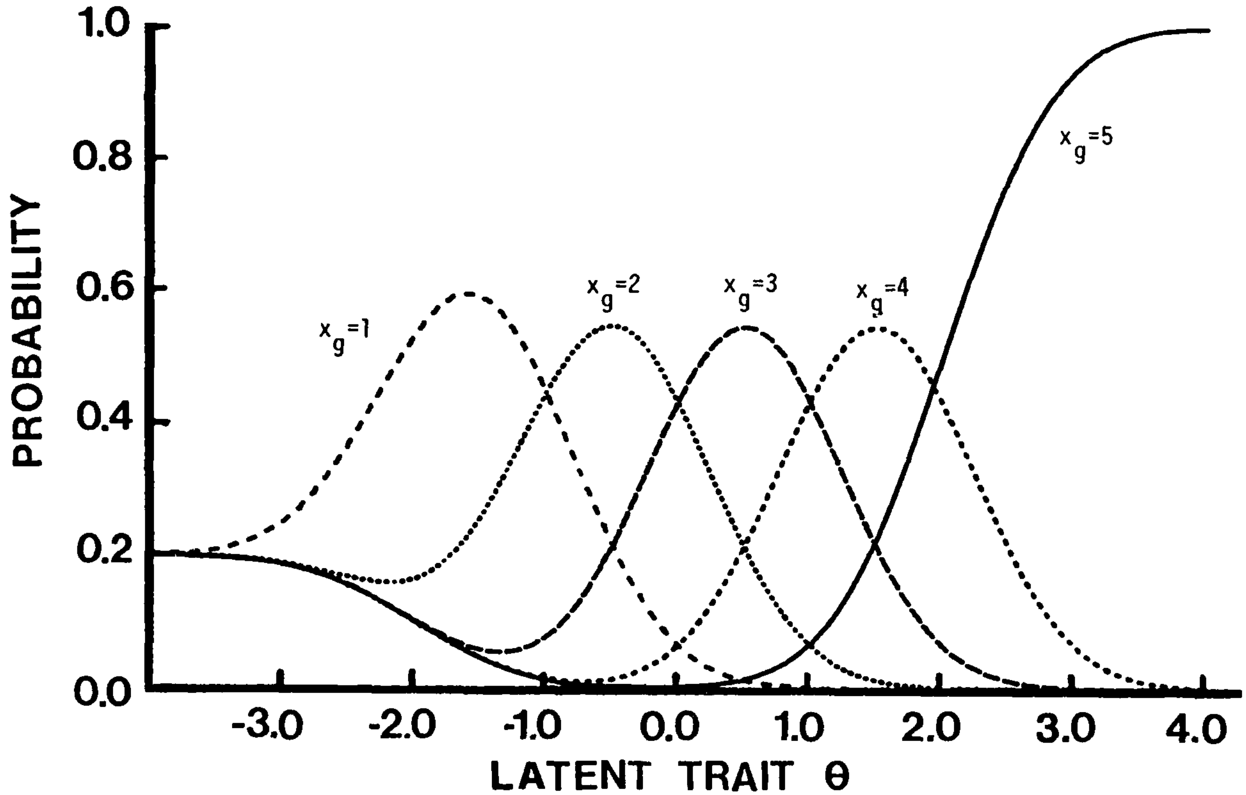


FIGURE 8-2

Operating Characteristics of Five Alternatives Following the Normal Ogive Model with Guessing Effect. The Parameters Are: $a_g = 1.5$, $b_1 = -2.0$, $b_2 = -1.0$, $b_3 = 0.0$, $b_4 = 1.0$ and $b_5 = 2.0$.

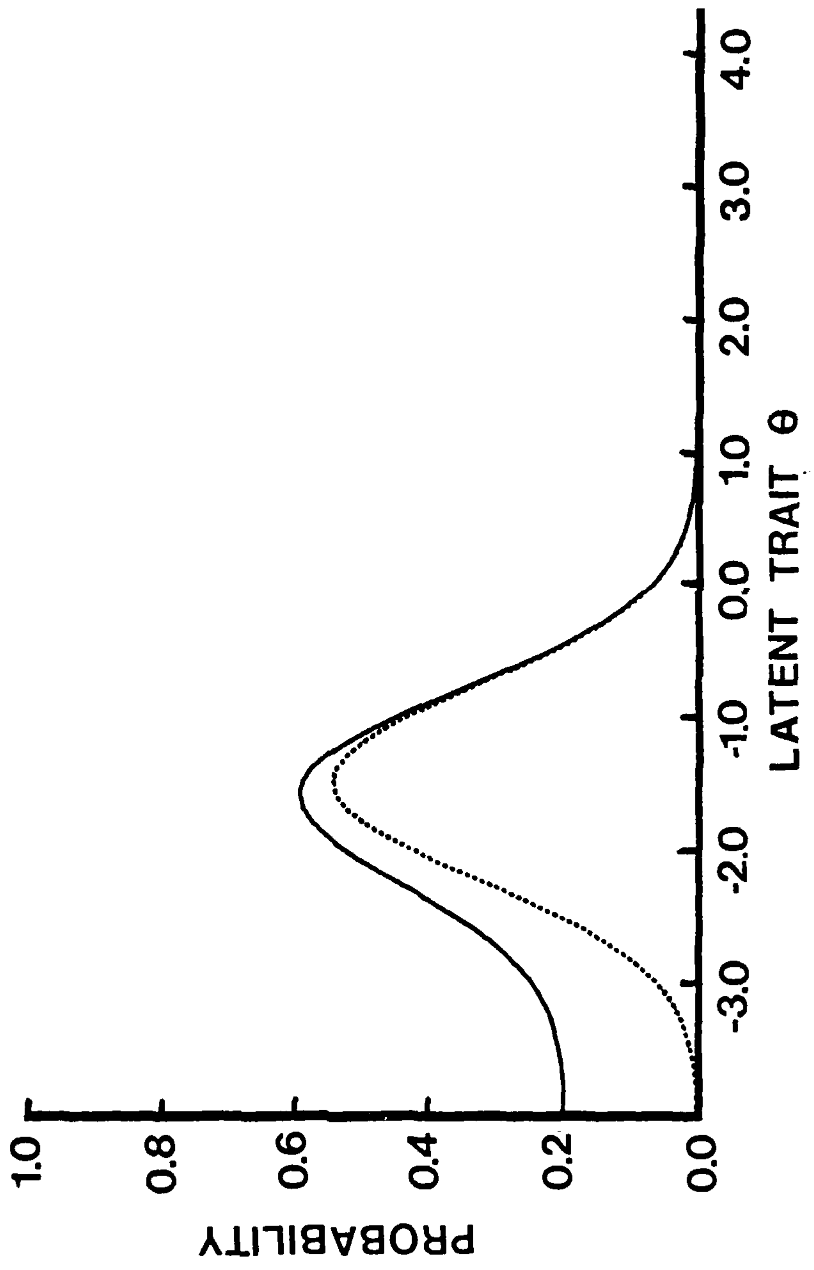


FIGURE 8-3

Comparison of the Two Operating Characteristics in the Normal Ogive Model.

$x_g = 1$.

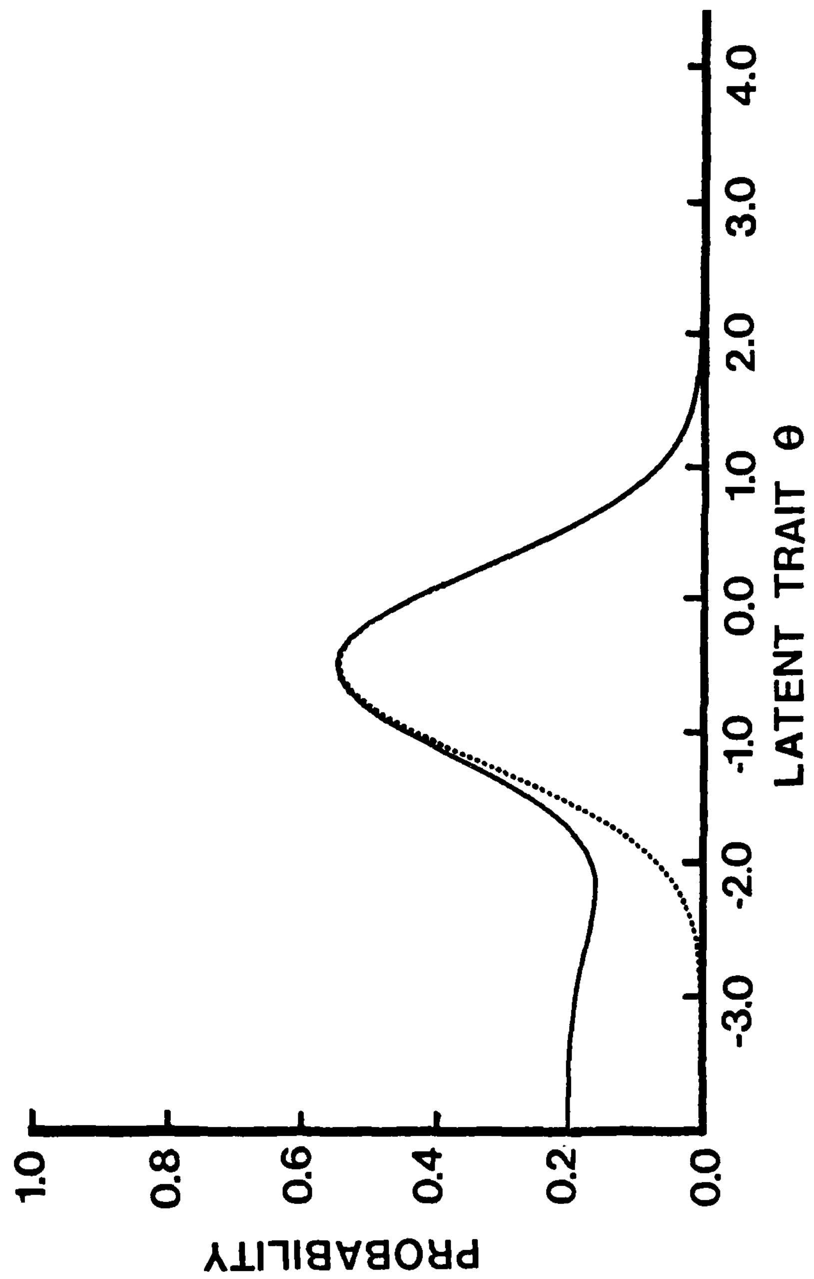


FIGURE 8-3 (Continued) $x_g = 2$.

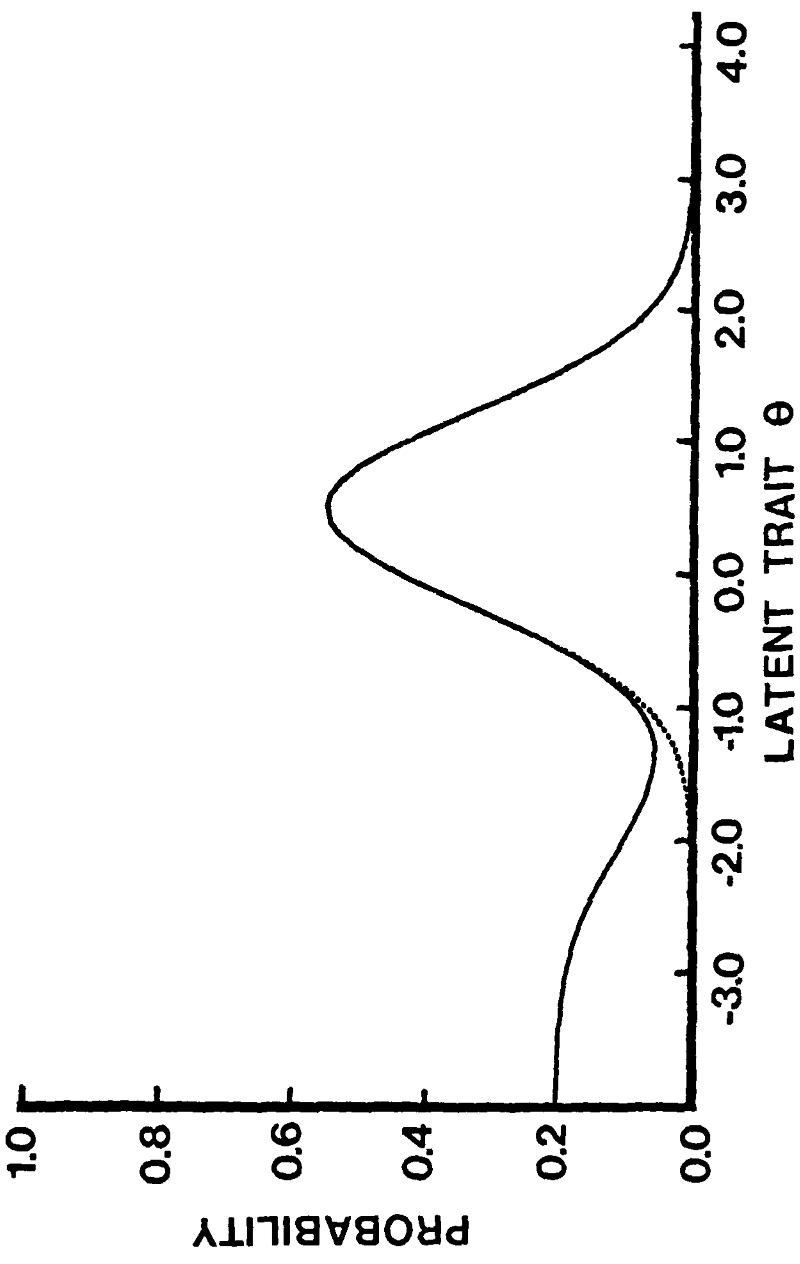


FIGURE 8-3 (Continued) $x_g = 3$.

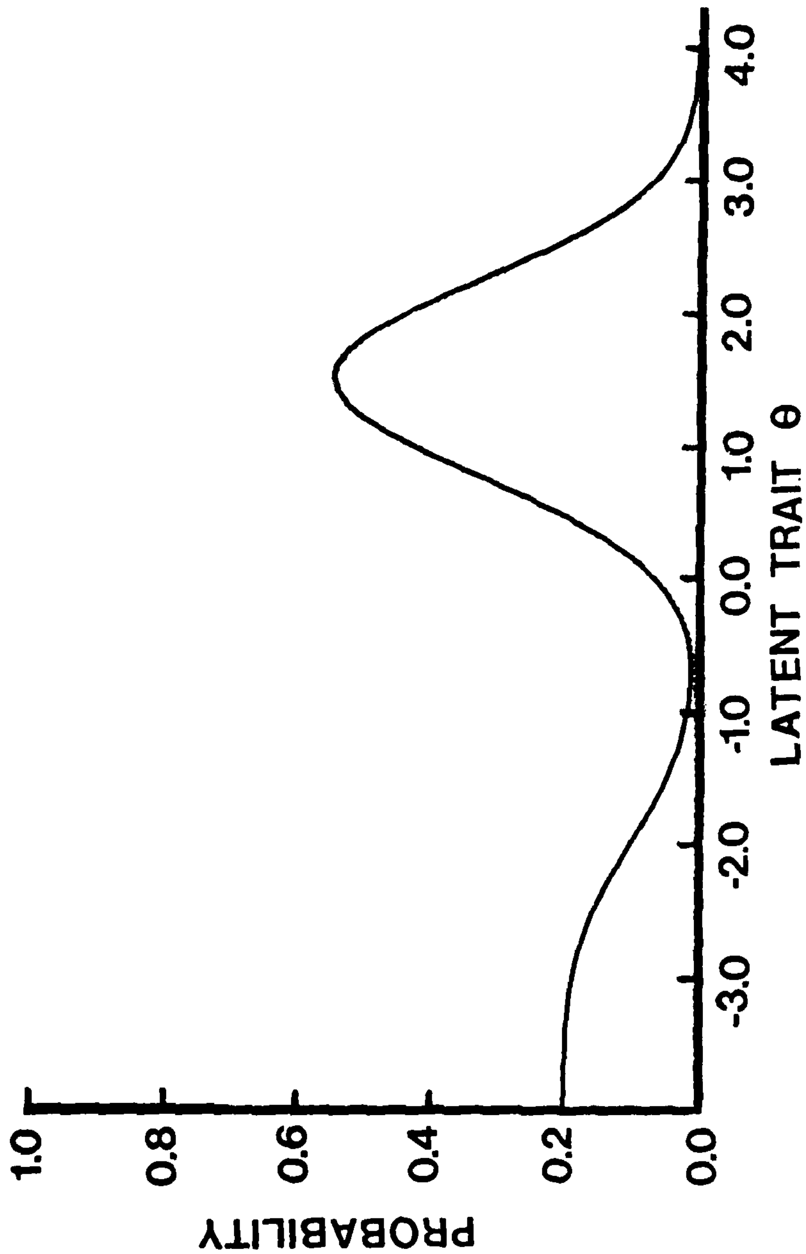


FIGURE 8-3 (Continued) $x_g = 4$.

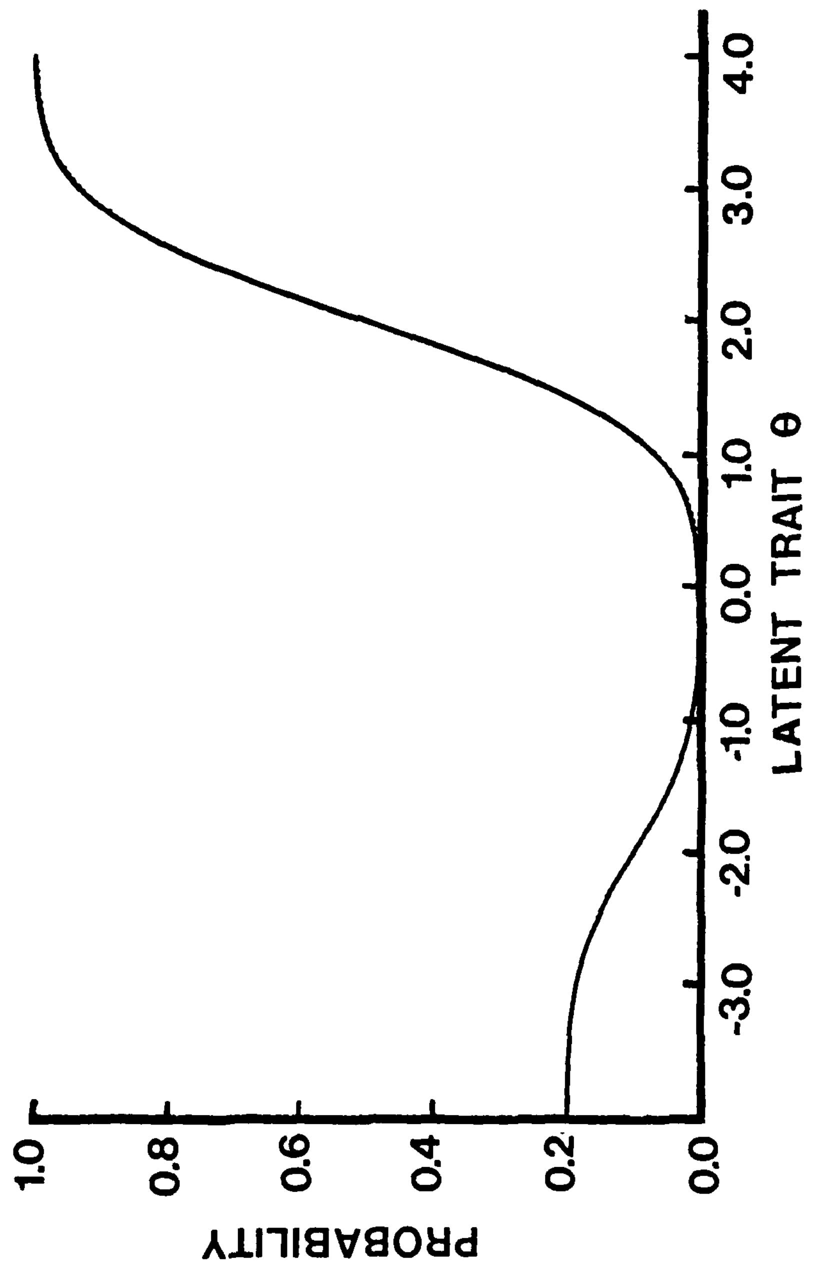


FIGURE 8-3 (Continued) $x_g = 5$.

this is the case with all the items in the test and the ability distribution of our examinees does not include lower levels of θ where these tails lie, we can approximate the operating characteristic of the correct answer by the normal ogive model on the dichotomous response level, and use the tetrachoric correlation coefficient and the logistic approximation and so on, just as Shiba and others did.

IX <u>Discussion and Conclusions</u>

We have introduced Sato's number of hypothetical, equivalent alternatives, and defined its modification, index k^* , as a measure of invalidating the three-parameter logistic, or normal ogive, model. We have also introduced Shiba's research on the measurement of vocabulary and the construction of a tailored test, using the information given by distractors. Various observations and discussion have been made concerning the three-parameter models and item distractors, the validation of mathematical models, and so forth. Finally, a new family of models for the multiple-choice item, which formulate both the operating characteristics of distractors and the effect of random guessing, has been proposed.

There is a tendency that researchers restrict their ideas within the tradition of their own culture. Thus they tend to accept whatever is familiar to them, what is fashionable among other researchers in their culture, and so on, without feeling the necessity of validating the ideas and mathematical models in relation with their specific data and psychological reality. The virtue of doubt can be obtained if they shift their attention to what is going on outside of their own culture and climate, and try to think what is really right.

Three-parameter models for the multiple-choice item have been too readily accepted among psychometricians and applied psychologists, and they have been using the models without trying to validate them. Unless we correct this wrong orientation, psychology will never make any progress, regardless of the fact

that more data are accumulated and more papers are published year by year. In the author's opinion, psychology has not yet established itself as a science, and we need to do that by putting ourselves in a right track of research. In so doing, the validation of mathematical models is certainly one of the most important things.

The departure from the tradition should also be made in the treatment of the multiple-choice item. Instead of trying to handle the multiple-choice item as a "blurred" substitute for the free-response item, we must make full use of its advantage, which the free-response item does not have. The operating characteristics of the distractors of the multiple-choice item will add more information about the examinee's ability level. We must set a criterion for the quality of multiple-choice items from this aspect also.

REFERENCES

[1] Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick; Statistical theories of mental test scores. Addison-Wesley, 1968, Chapters 17-20.

[2] Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 1972, pages 29-51.

[3] Goldman, S. Information theory. Prentice Hall, 1953.

[4] Indow, T. & F. Samejima. LIS measurement scale for non-verbal reasoning ability. Tokyo: Nippon Bunka Kagakusha, 1962. (in Japanese)

[5] Indow, T. & F. Samejima. On the results obtained by the absolute scaling model and the Lord model in the field of intelligence. Yokohama: Psychological Laboratory, Hiyoshi Campus, Keio University, 1966.

[6] Lawley, D. N. and A. E. Maxwell. Factor analysis as a statistical method. Butterworth, 1971.

[7] Lord, F. M. and M. R. Novick. Statistical theories of mental test scores. Addison-Wesley, 1968, Chapter 16.

[8] Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No. 17, 1969.

[9] Samejima, F. A general model for free-response data. Psychometrika Monograph, No. 18, 1972.

[10] Samejima, F. Homogeneous case of the continuous response model. Psychometrika, 1973, 38, pages 203-219.

[11] Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1, 1977a, pages 233-247.

[12] Samejima, F. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. Psychometrika, 42, 1977b, pages 163-191.

[13] Samejima, F. Estimation of the operating characteristics of item response categories I: Introduction to the Two-Parameter Beta Method. Office of Naval Research, Research Report 77-1, 1977c.

[14] Samejima, F. Estimation of the operating characteristics of item response categories II: Further development of the Two-Parameter Beta Method. Office of Naval Research, Research Report 78-1, 1978a.

REFERENCES (continued)

[15] Samejima, F. Estimation of the operating characteristics of item response categories III: The Normal Approach Method and the Pearson System Method. Office of Naval Research, Research Report 78-2, 1978b.

[16] Samejima, F. Estimation of the operating characteristics of item response categories IV: Comparison of the different methods. Office of Naval Research, Research Report 78-3, 1978c.

[17] Samejima, F. Estimation of the operating characteristics of item response categories V: Weighted Sum Procedure in the Conditional P.D.F. Approach. Office of Naval Research, Research Report 78-4, 1978d.

[18] Samejima, F. Estimation of the operating characteristics of item response categories VI: Proportioned Sum Procedure in the Conditional P.D.F. Approach. Office of Naval Research, Research Report 78-5, 1978e.

[19] Samejima, F. Estimation of the operating characteristics of item response categories VII: Bivariate P.D.F. Approach with Normal Approach Method. Office of Naval Research, Research Report 78-6, 1978f.

[20] Sato, T. Engineering techniques for analyzing educational data, No. 5. Central Research Institute, Nippon Electric Company, Ltd., 1977. (in Japanese)

[21] Shiba, S. Construction of a scale for acquistion of word meanings. Bulletin of Faculty of Education, University of Tokyo, 17, 1978. (in Japanese)

[22] Shiba, S., Y. Noguchi and T. Haebara. A stratified adaptive test of verbal ability. Japanese Journal of Educational Psychology, 16, 1978, pages 229-238. (in Japanese)

[23] Tatsuoka, M. M. Recent psychometric developments in Japan: engineers tackle educational measurement problems. ONR Tokyo Scientific Bulletin, 4 (1), 1979, pages 1-7.

APPENDIX I

TABLE A-1

Alternatives Selected by Five Hundred Hypothetical Examinees Following the Three Parameter Normal Ogive Model, for Each of the Five Hypothetical Items.
(1=A, 2=B, 3=C, 4=D and 5=E.)

Subject	Alternative				
	A	B	C	D	E
1	3	2	3	2	4
2	1	3	3	4	5
3	4	2	5	3	2
4	5	5	2	5	1
5	3	4	4	4	1
6	2	2	3	1	3
7	1	2	2	5	5
8	1	5	4	2	1
9	4	5	3	2	2
10	2	3	2	5	4
11	1	5	2	2	4
12	1	4	5	1	5
13	3	5	3	2	3
14	1	4	5	1	4
15	3	5	3	2	3
16	5	4	1	1	4
17	4	2	2	4	2
18	2	5	3	3	1
19	1	3	4	4	5
20	1	2	4	1	3
21	1	2	3	2	5
22	4	1	2	5	4
23	1	2	1	4	1
24	1	1	5	4	1
25	2	4	4	3	4
26	3	2	5	2	5
27	4	3	5	4	3
28	5	3	5	3	2
29	4	4	3	2	4
30	2	4	5	2	4
31	1	2	3	5	3
32	4	1	3	4	2
33	4	1	5	4	5
34	5	3	3	2	5
35	1	4	2	5	1
36	1	4	3	5	1
37	4	2	4	4	2
38	4	4	4	1	5
39	1	4	2	4	5
40	4	1	4	5	2
41	3	2	2	5	3
42	3	5	1	4	4
43	5	1	2	3	4
44	1	2	1	5	5
45	4	3	3	2	3
46	3	2	1	4	4
47	3	4	4	1	2
48	2	1	4	1	2
49	2	1	4	2	5
50	2	2	3	4	3
51	3	1	4	2	4
52	4	1	3	1	1
53	2	5	4	3	2
54	3	1	2	2	5
55	3	5	3	5	3
56	2	4	5	2	1
57	2	4	1	4	1
58	5	1	5	2	4
59	1	2	3	5	5
60	1	3	5	4	1
61	2	2	1	1	2
62	3	5	4	4	5
63	5	1	4	2	5
64	1	1	1	4	2
65	5	5	5	4	1
66	2	3	4	4	3
67	5	1	3	5	5
68	1	1	1	5	5
69	2	2	1	1	1
70	5	2	2	5	3
71	1	4	1	4	5
72	5	1	2	3	2
73	5	2	4	2	2
74	5	1	1	4	3
75	2	3	4	1	4
76	4	1	5	1	1
77	4	3	2	3	4
78	4	2	2	2	4
79	3	1	2	5	3
80	1	3	5	4	1
81	5	4	4	4	4
82	4	1	5	1	5
83	3	3	5	4	1
84	3	2	1	5	4
85	2	5	5	1	4
86	1	2	4	3	1
87	4	5	2	4	1
88	1	4	4	1	4
89	3	2	2	3	3
90	1	4	1	1	2
91	5	4	2	5	2
92	2	4	2	5	5
93	1	1	1	4	1
94	3	3	5	1	2
95	3	3	4	1	2
96	2	5	1	3	1
97	5	1	3	2	2
98	3	1	5	4	5
99	2	4	1	2	3
100	4	4	2	5	1
101	2	1	2	5	3
102	5	2	3	3	3
103	3	2	1	3	3
104	3	2	3	5	5
105	3	1	5	3	5
106	4	2	3	3	1
107	2	3	3	5	3
108	3	4	3	1	5
109	1	1	4	5	3
110	1	5	3	3	3
111	5	4	1	2	5
112	1	4	4	5	5
113	2	4	1	3	4
114	5	4	5	4	1
115	1	5	5	2	3
116	1	1	5	1	4
117	5	5	4	3	5
118	5	5	3	1	5
119	1	1	3	4	1
120	5	4	3	1	3
121	1	5	1	4	5
122	2	5	1	5	4
123	1	3	2	4	1
124	5	3	3	1	1
125	5	4	2	2	4
126	5	2	5	1	1
127	4	4	2	4	5
128	1	5	1	4	1
129	1	5	1	2	4
130	3	1	5	2	3
131	1	1	3	4	4
132	3	3	2	1	1
133	1	4	1	1	3
134	3	4	5	1	1
135	5	5	2	3	3
136	3	4	4	3	3
137	1	1	1	4	1
138	5	3	1	4	5
139	4	2	4	4	2
140	3	5	3	3	3
141	1	2	1	3	5
142	1	5	1	2	1
143	1	2	4	5	4
144	1	1	4	4	4
145	3	2	1	5	3
146	3	4	4	1	3
147	3	4	3	4	4
148	4	2	4	4	1
149	3	1	1	1	4
150	1	5	4	1	3
151	3	5	5	2	2
152	2	1	4	4	3
153	1	1	1	4	2
154	4	3	5	4	1
155	1	4	2	3	2
156	4	4	5	2	3
157	4	1	2	2	5
158	2	5	5	4	1
159	1	4	4	1	4
160	3	2	4	3	1
161	4	1	5	5	1
162	1	5	5	2	1
163	3	3	1	5	3
164	4	5	4	3	5
165	4	2	1	3	1
166	2	3	4	1	5
167	5	5	3	5	2
168	1	5	4	3	4
169	1	1	3	4	4
170	5	3	4	4	3
171	1	5	1	4	1
172	2	1	2	2	1
173	5	5	2	1	1
174	1	5	1	3	5
175	1	1	1	4	4
176	4	1	1	1	5
177	1	2	3	5	2
178	1	3	1	2	2
179	1	1	1	4	2
180	1	1	2	1	3
181	3	2	4	2	5
182	3	2	1	3	4
183	4	3	5	3	3
184	5	2	1	2	4
185	3	1	2	1	4
186	4	3	1	1	5
187	1	2	1	2	4
188	1	2	4	1	2
189	5	1	2	2	4
190	4	1	5	3	4
191	1	1	4	1	1
192	4	5	1	1	4
193	2	5	5	2	4
194	4	3	4	5	2
195	1	1	4	2	2
196	1	1	5	5	4
197	3	1	3	3	5
198	1	5	3	5	4
199	1	1	2	5	1
200	1	4	1	2	3

TABLE A-1 (Continued)

Subject	Alternative A	B	C	D	E
201	1	4	3	4	2
202	2	5	3	1	5
203	1	1	4	2	5
204	1	4	3	4	5
205	3	4	4	1	4
206	1	2	4	1	2
207	5	1	2	1	2
208	1	1	1	1	2
209	1	1	4	3	2
210	5	3	4	2	2
211	4	4	3	2	3
212	4	2	2	5	3
213	4	4	1	1	3
214	1	3	2	4	5
215	5	1	2	1	5
216	5	3	3	3	1
217	5	3	1	2	5
218	1	3	1	4	4
219	2	2	3	4	1
220	3	1	1	4	5
221	1	1	1	4	5
222	1	1	4	1	1
223	1	4	1	3	2
224	4	1	5	3	1
225	1	1	5	4	1
226	1	2	4	1	1
227	4	4	3	1	4
228	1	1	5	1	2
229	1	1	1	1	3
230	5	1	2	1	2
231	1	3	1	4	3
232	5	4	1	2	1
233	1	2	1	2	1
234	1	3	1	1	3
235	1	1	3	1	2
236	3	5	1	1	2
237	1	4	3	1	5
238	1	4	3	4	4
239	1	3	1	2	3
240	2	4	1	1	1
241	1	1	1	4	4
242	1	2	1	1	1
243	4	1	1	1	1
244	1	1	2	4	4
245	1	1	1	5	2
246	1	1	3	1	5
247	1	1	5	1	1
248	1	1	5	1	1
249	1	5	1	1	1
250	3	1	1	1	1
251	1	4	4	1	1
252	2	1	5	4	1
253	1	1	1	1	5
254	4	1	1	5	4
255	5	3	1	5	1
256	4	1	2	1	1
257	2	1	2	1	3
258	1	1	1	2	3
259	1	5	1	1	1
260	1	4	1	1	3
261	1	3	1	1	2
262	1	1	3	5	1
263	1	1	1	1	1
264	1	1	5	1	2
265	1	1	5	1	2
266	1	1	1	4	1
267	1	5	1	1	1
268	2	1	1	1	3
269	5	3	4	1	1
270	4	4	1	1	1
271	1	1	1	2	1
272	4	1	3	1	1
273	2	1	1	1	1
274	4	1	1	3	1
275	1	2	1	1	1
276	1	1	2	2	1
277	5	5	1	1	1
278	1	1	1	1	1
279	1	1	1	1	1
280	1	1	1	1	1
281	3	1	2	1	1
282	1	1	1	1	1
283	1	1	1	1	1
284	1	1	1	1	1
285	1	1	1	1	1
286	3	2	1	1	5
287	2	1	1	1	1
288	1	1	1	1	1
289	1	1	1	1	1
290	1	1	1	1	1
291	1	1	1	1	1
292	1	1	1	1	1
293	5	1	1	1	1
294	1	5	1	1	2
295	1	1	1	1	1
296	1	1	1	1	1
297	5	1	3	1	1
298	5	2	1	1	1
299	4	1	1	2	1
300	1	1	1	3	1
301	5	1	1	1	1
302	1	1	1	1	1
303	3	1	1	1	1
304	1	1	1	4	1
305	2	1	1	1	1
306	3	1	5	1	1
307	1	1	1	1	1
308	1	1	1	1	1
309	4	1	1	1	1
310	1	1	1	1	1
311	1	1	1	1	1
312	5	1	2	1	1
313	1	1	1	1	1
314	1	5	1	1	1
315	1	1	1	1	1
316	2	3	1	1	1
317	5	1	1	1	1
318	1	1	1	1	1
319	1	1	1	1	1
320	1	1	1	1	1
321	1	1	1	1	1
322	1	1	2	1	1
323	1	1	1	1	1
324	1	1	1	1	1
325	1	1	1	1	1
326	1	1	1	1	1
327	1	1	1	1	1
328	1	1	1	1	1
329	5	1	1	1	1
330	1	1	1	1	1
331	1	1	1	1	1
332	1	1	1	1	1
333	1	1	1	1	1
334	1	1	1	1	1
335	1	1	1	3	1
336	3	1	1	1	1
337	1	1	1	1	1
338	1	1	1	1	1
339	1	1	1	1	1
340	1	1	1	1	1
341	1	1	1	1	1
342	2	1	1	1	1
343	1	1	1	1	1
344	1	1	1	1	1
345	1	1	3	1	1
346	3	1	1	1	1
347	1	1	1	1	1
348	1	1	1	1	1
349	1	1	1	1	1
350	1	1	1	1	1
351	1	1	1	1	1
352	1	1	1	1	1
353	1	1	1	1	1
354	1	1	1	1	1
355	1	1	1	1	1
356	1	1	1	1	1
357	1	1	1	1	1
358	1	1	1	1	1
359	2	1	1	1	1
360	3	1	4	1	1
361	1	1	1	1	1
362	1	1	1	1	1
363	1	1	1	1	1
364	1	1	1	1	1
365	1	1	1	1	1
366	1	1	1	1	1
367	1	1	1	1	1
368	3	1	1	1	1
369	1	1	1	1	1
370	1	1	1	1	1
371	1	1	1	1	1
372	1	1	1	1	1
373	1	1	1	1	1
374	1	1	1	1	1
375	1	1	1	1	1
376	1	1	1	1	1
377	1	1	1	1	1
378	1	1	1	1	1
379	5	1	1	1	1
380	1	1	1	1	1
381	1	1	1	1	1
382	1	1	1	1	1
383	1	1	1	1	1
384	2	1	1	1	1
385	1	1	1	1	1
386	1	1	1	1	1
387	1	1	1	1	1
388	1	1	1	1	1
389	2	1	1	1	1
390	1	1	1	1	1
391	1	1	1	1	1
392	1	1	1	1	1
393	1	1	1	1	1
394	4	1	1	1	1
395	1	1	1	1	1
396	1	1	1	1	1
397	1	1	1	1	1
398	1	1	1	1	1
399	1	1	1	1	1
400	1	1	1	1	1

TABLE A-1 (Continued)

Subject	Alternative A	B	C	D	E
401	1	1	1	1	1
402	1	1	1	1	1
403	1	1	1	1	1
404	1	1	1	1	1
405	1	1	1	1	1
406	1	1	1	1	1
407	1	1	1	1	1
408	1	1	1	1	1
409	3	1	1	1	1
410	4	1	1	1	1
411	1	1	1	1	1
412	1	2	1	1	1
413	1	1	1	1	1
414	1	1	1	1	1
415	1	1	1	1	1
416	1	1	1	1	1
417	1	1	1	1	1
418	1	1	1	1	1
419	1	1	1	1	1
420	1	1	1	1	1
421	2	1	1	1	1
422	1	1	1	1	1
423	1	1	1	1	1
424	1	1	1	1	1
425	1	1	1	1	1
426	1	1	1	1	1
427	1	1	1	1	1
428	1	1	1	1	1
429	1	1	1	1	1
430	1	1	1	1	1
431	1	1	1	1	1
432	1	1	1	1	1
433	1	1	1	1	1
434	1	1	1	1	1
435	1	1	1	1	1
436	1	1	1	1	1
437	1	1	1	1	1
438	1	1	1	1	1
439	1	1	1	1	1
440	1	1	1	1	1
441	1	1	1	1	1
442	1	1	1	1	1
443	1	1	1	1	1
444	1	1	1	1	1
445	1	1	1	1	1
446	1	1	1	1	1
447	1	1	1	1	1
448	1	1	1	1	1
449	1	1	1	1	1
450	1	1	1	1	1

Subject	Alternative A	B	C	D	E
451	1	1	1	1	1
452	1	1	1	1	1
453	1	1	1	1	1
454	1	1	1	1	1
455	1	1	1	1	1
456	1	1	1	1	1
457	1	1	1	1	1
458	1	1	1	1	1
459	1	1	1	1	1
460	1	1	1	1	1
461	1	1	1	1	1
462	1	1	1	1	1
463	1	1	1	1	1
464	1	1	1	1	1
465	1	1	1	1	1
466	1	1	1	1	1
467	1	1	1	1	1
468	1	1	1	1	1
469	1	1	1	1	1
470	1	1	1	1	1
471	1	1	1	1	1
472	1	1	1	1	1
473	1	1	1	1	1
474	1	1	1	1	1
475	1	1	1	1	1
476	1	1	1	1	1
477	1	1	1	1	1
478	1	1	1	1	1
479	1	1	1	1	1
480	1	1	1	1	1
481	1	1	1	1	1
482	1	1	1	1	1
483	1	1	1	1	1
484	1	1	1	1	1
485	1	1	1	1	1
486	1	1	1	1	1
487	1	1	1	1	1
488	1	1	1	1	1
489	1	1	1	1	1
490	1	1	1	1	1
491	1	1	1	1	1
492	1	1	1	1	1
493	1	1	1	1	1
494	1	1	1	1	1
495	1	1	1	1	1
496	1	1	1	1	1
497	1	1	1	1	1
498	1	1	1	1	1
499	1	1	1	1	1
500	1	1	1	1	1

AFPENDIX II

TABLE A-2

Frequency Ratio, P_j , of Each of the Five Alternatives and the Estimated Probability, P_R^* , for the Correct Answer with Which the Examinee Selects the Correct Answer by Random Guessing at the Maximum, for Each of the Nineteen Vocabulary Items. Junior High School, Grade 1, for Test J1.

Item	P_j and P_R^*	Alternative 1	2	3	4	5
37 (1)	RELATIVE FREQUENCY	0.50175	0.08741	0.10315	0.10315	0.20455
	MODIFIED REL.FREQ.	0.13271				
39 (3)	RELATIVE FREQUENCY	0.16192	0.20463	0.20996	0.09075	0.33274
	MODIFIED REL.FREQ.					0.17450
40 (4)	RELATIVE FREQUENCY	0.10471	0.24607	0.15707	0.47644	0.01571
	MODIFIED REL.FREQ.				0.16692	
41 (5)	RELATIVE FREQUENCY	0.09250	0.03490	0.04014	0.14834	0.68412
	MODIFIED REL.FREQ.					0.09324
42 (6)	RELATIVE FREQUENCY	0.43333	0.03684	0.21228	0.14737	0.17018
	MODIFIED REL.FREQ.	0.16122				
43 (7)	RELATIVE FREQUENCY	0.04545	0.53846	0.17133	0.11713	0.12762
	MODIFIED REL.FREQ.		0.12583			
44 (8)	RELATIVE FREQUENCY	0.20914	0.11775	0.02460	0.58524	0.06327
	MODIFIED REL.FREQ.				0.13040	
45 (9)	RELATIVE FREQUENCY	0.08979	0.04401	0.21655	0.60915	0.04049
	MODIFIED REL.FREQ.				0.12427	
46(10)	RELATIVE FREQUENCY	0.52113	0.08099	0.07746	0.28873	0.03169
	MODIFIED REL.FREQ.	0.16263				
47(11)	RELATIVE FREQUENCY	0.12127	0.39367	0.27768	0.09315	0.11424
	MODIFIED REL.FREQ.		0.16628			
48(12)	RELATIVE FREQUENCY	0.14362	0.17730	0.10284	0.11879	0.45745
	MODIFIED REL.FREQ.					0.13854
49(13)	RELATIVE FREQUENCY	0.20070	0.05410	0.07330	0.12216	0.54974
	MODIFIED REL.FREQ.					0.12718
50(14)	RELATIVE FREQUENCY	0.53940	0.08056	0.06130	0.15061	0.16813
	MODIFIED REL.FREQ.	0.12468				
51(15)	RELATIVE FREQUENCY	0.15892	0.14643	0.13393	0.20179	0.35893
	MODIFIED REL.FREQ.					0.16225
52(16)	RELATIVE FREQUENCY	0.20531	0.41593	0.14159	0.06018	0.17699
	MODIFIED REL.FREQ.		0.15807			
53(17)	RELATIVE FREQUENCY	0.28497	0.08916	0.05944	0.25000	0.31643
	MODIFIED REL.FREQ.				0.22911	
54(18)	RELATIVE FREQUENCY	0.32442	0.19786	0.17825	0.25312	0.04635
	MODIFIED REL.FREQ.	0.19109				
55(19)	RELATIVE FREQUENCY	0.04729	0.12609	0.55517	0.05079	0.22067
	MODIFIED REL.FREQ.					0.32187
56(20)	RELATIVE FREQUENCY	0.18182	0.17657	0.20105	0.24650	0.19406
	MODIFIED REL.FREQ.				0.18862	

TABLE A-2 (Continued): Junior High School, Grade 2, for Test J1.

Item	P_j and P_R^*	Alternative 1	2	3	4	5
37 (1)	RELATIVE FREQUENCY	0.59121	0.08571	0.08571	0.08132	0.15604
	MODIFIED REL.FREQ.	0.10662				
39 (3)	RELATIVE FREQUENCY	0.12222	0.21556	0.18222	0.11111	0.36889
	MODIFIED REL.FREQ.					0.16372
40 (4)	RELATIVE FREQUENCY	0.13319	0.20742	0.09825	0.53057	0.03057
	MODIFIED REL.FREQ.				0.13810	
41 (5)	RELATIVE FREQUENCY	0.05857	0.02620	0.04121	0.10195	0.77007
	MODIFIED REL.FREQ.					0.06429
42 (6)	RELATIVE FREQUENCY	0.56236	0.03063	0.14880	0.14661	0.11160
	MODIFIED REL.FREQ.	0.12318				
43 (7)	RELATIVE FREQUENCY	0.02183	0.68122	0.11572	0.10044	0.08079
	MODIFIED REL.FREQ.		0.09013			
44 (8)	RELATIVE FREQUENCY	0.14254	0.11842	0.03289	0.63816	0.06798
	MODIFIED REL.FREQ.				0.10216	
45 (9)	RELATIVE FREQUENCY	0.06536	0.05011	0.17211	0.65142	0.06100
	MODIFIED REL.FREQ.				0.10025	
46(10)	RELATIVE FREQUENCY	0.67102	0.05447	0.06318	0.19608	0.01525
	MODIFIED REL.FREQ.	0.11336				
47(11)	RELATIVE FREQUENCY	0.09368	0.65795	0.11765	0.06536	0.06536
	MODIFIED REL.FREQ.		0.08629			
48(12)	RELATIVE FREQUENCY	0.12308	0.18681	0.08352	0.09890	0.50769
	MODIFIED REL.FREQ.					0.12921
49(13)	RELATIVE FREQUENCY	0.20870	0.03478	0.04130	0.07609	0.63913
	MODIFIED REL.FREQ.					0.11792
50(14)	RELATIVE FREQUENCY	0.59121	0.04176	0.04396	0.18462	0.13846
	MODIFIED REL.FREQ.	0.12331				
51(15)	RELATIVE FREQUENCY	0.16964	0.10491	0.12277	0.15179	0.45089
	MODIFIED REL.FREQ.					0.13961
52(16)	RELATIVE FREQUENCY	0.13363	0.53229	0.15813	0.04677	0.12918
	MODIFIED REL.FREQ.		0.12617			
53(17)	RELATIVE FREQUENCY	0.20479	0.07407	0.05664	0.39216	0.27233
	MODIFIED REL.FREQ.				0.18236	
54(18)	RELATIVE FREQUENCY	0.39246	0.16630	0.18182	0.23947	0.01996
	MODIFIED REL.FREQ.	0.18393				
55(19)	RELATIVE FREQUENCY	0.04367	0.09625	0.37991	0.03930	0.43886
	MODIFIED REL.FREQ.					0.21613
56(20)	RELATIVE FREQUENCY	0.15385	0.21978	0.17582	0.28132	0.16923
	MODIFIED REL.FREQ.				0.18130	

TABLE A-2 (Continued): Junior High School, Grade 2, for Test J2.

Item	P_j and P_R^*	Alternative 1	2	3	4	5
1(37)	RELATIVE FREQUENCY	0.65611	0.04977	0.08597	0.04977	0.15837
	MODIFIED REL.FREQ.	0.09724				
3(39)	RELATIVE FREQUENCY	0.12844	0.20642	0.19266	0.07339	0.39908
	MODIFIED REL.FREQ.					0.16079
4(40)	RELATIVE FREQUENCY	0.23077	0.13575	0.10407	0.52036	0.00905
	MODIFIED REL.FREQ.				0.15717	
5(41)	RELATIVE FREQUENCY	0.06335	0.00905	0.04525	0.08145	0.80090
	MODIFIED REL.FREQ.					0.05953
6(42)	RELATIVE FREQUENCY	0.56818	0.02727	0.14545	0.14545	0.11364
	MODIFIED REL.FREQ.	0.12263				
7(43)	RELATIVE FREQUENCY	0.00452	0.69231	0.10860	0.13575	0.05882
	MODIFIED REL.FREQ.		0.10171			
8(44)	RELATIVE FREQUENCY	0.16742	0.05430	0.02715	0.70588	0.04525
	MODIFIED REL.FREQ.				0.09401	
9(45)	RELATIVE FREQUENCY	0.04545	0.04091	0.09545	0.78182	0.03636
	MODIFIED REL.FREQ.				0.05940	
10(46)	RELATIVE FREQUENCY	0.63801	0.04977	0.08145	0.21267	0.01810
	MODIFIED REL.FREQ.	0.12408				
11(47)	RELATIVE FREQUENCY	0.10000	0.65000	0.11818	0.03182	0.10000
	MODIFIED REL.FREQ.		0.09534			
12(48)	RELATIVE FREQUENCY	0.10407	0.10407	0.04525	0.08145	0.66516
	MODIFIED REL.FREQ.					0.08761
13(49)	RELATIVE FREQUENCY	0.22624	0.02715	0.09502	0.05882	0.59276
	MODIFIED REL.FREQ.					0.13206
14(50)	RELATIVE FREQUENCY	0.60909	0.05909	0.04091	0.15455	0.13636
	MODIFIED REL.FREQ.	0.11132				
15(51)	RELATIVE FREQUENCY	0.15385	0.08145	0.11765	0.16290	0.48416
	MODIFIED REL.FREQ.					0.13327
16(52)	RELATIVE FREQUENCY	0.16590	0.40092	0.24885	0.05069	0.13364
	MODIFIED REL.FREQ.		0.16923			
17(53)	RELATIVE FREQUENCY	0.23529	0.09502	0.02262	0.38462	0.26244
	MODIFIED REL.FREQ.				0.19663	
18(54)	RELATIVE FREQUENCY	0.49772	0.15068	0.14155	0.17808	0.03196
	MODIFIED REL.FREQ.	0.14233				
19(55)	RELATIVE FREQUENCY	0.02262	0.17195	0.39819	0.03167	0.37557
	MODIFIED REL.FREQ.					0.25048
20(56)	RELATIVE FREQUENCY	0.20814	0.17195	0.20814	0.30317	0.10860
	MODIFIED REL.FREQ.				0.17941	

TABLE A-2 (Continued): Junior High School, Grade 3, for Test J2.

Item	P_j and P_R^*	Alternative 1	2	3	4	5
1(37)	RELATIVE FREQUENCY	0.76091	0.05236	0.04363	0.03316	0.10995
	MODIFIED REL.FREQ.	0.06686				
3(39)	RELATIVE FREQUENCY	0.09524	0.16402	0.17108	0.11111	0.45855
	MODIFIED REL.FREQ.					0.13946
4(40)	RELATIVE FREQUENCY	0.14510	0.13462	0.06643	0.63287	0.02098
	MODIFIED REL.FREQ.				0.10973	
5(41)	RELATIVE FREQUENCY	0.05226	0.01220	0.03310	0.07491	0.82753
	MODIFIED REL.FREQ.					0.05051
6(42)	RELATIVE FREQUENCY	0.65317	0.01232	0.16901	0.08627	0.07923
	MODIFIED REL.FREQ.	0.10957				
7(43)	RELATIVE FREQUENCY	0.01754	0.77368	0.07895	0.08772	0.04211
	MODIFIED REL.FREQ.		0.06511			
8(44)	RELATIVE FREQUENCY	0.09632	0.05429	0.02452	0.80035	0.02452
	MODIFIED REL.FREQ.				0.05889	
9(45)	RELATIVE FREQUENCY	0.05614	0.08246	0.11754	0.71404	0.02982
	MODIFIED REL.FREQ.				0.07956	
10(46)	RELATIVE FREQUENCY	0.82837	0.02627	0.04904	0.08932	0.00701
	MODIFIED REL.FREQ.	0.05624				
11(47)	RELATIVE FREQUENCY	0.06294	0.68007	0.12063	0.06119	0.07517
	MODIFIED REL.FREQ.		0.08341			
12(48)	RELATIVE FREQUENCY	0.05769	0.15210	0.01748	0.06119	0.71154
	MODIFIED REL.FREQ.					0.09059
13(49)	RELATIVE FREQUENCY	0.18674	0.02792	0.05061	0.02443	0.71030
	MODIFIED REL.FREQ.					0.10427
14(50)	RELATIVE FREQUENCY	0.67776	0.03853	0.02102	0.14711	0.11559
	MODIFIED REL.FREQ.	0.10125				
15(51)	RELATIVE FREQUENCY	0.11943	0.04813	0.14082	0.12299	0.56863
	MODIFIED REL.FREQ.					0.11483
16(52)	RELATIVE FREQUENCY	0.10193	0.63972	0.13181	0.02460	0.10193
	MODIFIED REL.FREQ.		0.10162			
17(53)	RELATIVE FREQUENCY	0.19056	0.07343	0.05245	0.45280	0.23077
	MODIFIED REL.FREQ.				0.16063	
18(54)	RELATIVE FREQUENCY	0.52035	0.11504	0.14159	0.20354	0.01947
	MODIFIED REL.FREQ.	0.14498				
19(55)	RELATIVE FREQUENCY	0.02622	0.15210	0.23776	0.01573	0.56818
	MODIFIED REL.FREQ.					0.16095
20(56)	RELATIVE FREQUENCY	0.13589	0.12892	0.16202	0.40941	0.16376
	MODIFIED REL.FREQ.				0.14846	

APPENDIX III

RESEARCHERS OF THE UNIVERSITY OF TOYKO

Dr. Sukeyori Shiba
1-8-3 Minami-Ogikubo, Suginami-ku
Tokyo 167, Japan

Mr. Yukihiro Noguchi
1208-7 Kamoi, Midori-ku
Yokohama 226, Japan

Mr. Tomokazu Haebara
Lindquist Center for Measurement
The University of Iowa
Iowa City, Iowa 52242
U.S.A.

TECHNOLOGICAL RESEARCHERS IN EDUCATIONAL MEASUREMENT PROBLEMS

1. Educational Technology Group of IECE

Dr. Takahiro Sato (Representative) Phone: (044) 855-1111 ext. 2296
Application Research Laboratory
Central Research Laboratories
Nippon Electric Co., Ltd.
4-1-1 Miyazaki, Takatsu-ku
Kawasaki 213, Japan

Dr. Hiroshi Ikeda Phone: (03) 726-1111
Educational Technology Center
Tokyo Institute of Technology
2-12-1 Ohokayama, Meguro-ku
Tokyo 152, Japan

Dr. Hideo Fujiwara Phone: (06) 877-5111
Department of Electronic Engineering
Faculty of Engineering
Osaka University
Yamadakami, Suita-shi
Osaka 565, Japan

Dr. Keizo Nagaoka Phone: (078) 881-1212
Educational Technology Center
Department of Education
Kohbe University
3-11 Tsurukabuto, Nada-ku
Kohbe 657, Japan

Dr. Yoneo Yamamoto Phone: (0886) 23-2311
Department of Information Science
Faculty of Engineering
Tokushima University
2-1 Minamijosanjima-cho
Tokushima 770, Japan

(as of Jan. 4, 1979 -- University of Illinois, CERL)

Mr. Makoto Takeya Phone: (044) 855-1111
Application Research Laboratory
Central Research Laboratories
Nippon Electric Co., Ltd.,
4-1-1 Miyazaki, Takatsu-ku
Kawasaki 213, Japan

Mr. Masahiko Kurata — Phone: (044) 855-1111
Application Research Laboratory
Central Research Laboratories
Nippon Electric Co., Ltd.,
4-1-1 Miyazaki, Takatsu-ku
Kawasaki 213, Japan

Mr. Yasuhiro Morimoto — Phone: (044) 855-1111
Application Research Laboratory
Central Research Laboratories
Nippon Electric Co., Ltd.,
4-1-1 Miyazaki, Takatsu-ku
Kawasaki 213, Japan

Mr. Hiroyasu Chimura — Phone: (044) 855-1111
Application Research Laboratory
Central Research Laboratories
Nippon Electric Co., Ltd.,
4-1-1 Miyazaki, Takatsu-ku
Kawasaki 213, Japan

2. <u>Others</u>

Dr. Moriya Oda — Phone: (052) 781-5111
Research Institute of Cybernetics
Department of Engineering
Nagoya University
Furocho, Chikusa-ku
Nagoya 464, Japan

Dr. Masahi Ishiketa — Phone: (0720) 22-2161
Department of Engineering
Osaka Denki-Tsushin University
18-8 Hatsu-cho, Neyagawa-shi
Osaka-fu 572, Japan

Mr. Haruo Nishinosono — Phone: (075) 641-9281
Educational Technology Center
Kyoto University of Education
1 Fukakusa-Fujinomori-cho, Fushimi-ku
Kyoto 612, Japan

Dr. Haruo Sunouchi — Phone: (03) 209-3211
Department of Science and Engineering
Waseda University
3-4-1 Ohkubo, Shinjuku-ku
Tokyo 160, Japan

Mr. Hajime Yamashita — Phone: (03) 203-4141
Department of Politics and Economics
Waseda University
1-6-1 Nishi-Waseda, Shinjuku-ku
Tokyo 160, Japan

Mr. Masahiro Yokoi — Phone: (0427) 32-9111
Department of Engineering
Tamagawa University
6-1-1 Tamagawa-Gakuen, Machida-shi
Tokyo 194, Japan

Mr. Takeshi Kikukawa — Phone: (0463) 58-1211
Department of Communication Engineering
Faculty of Engineering
Tokai University
1117 Kitakaname, Hiratsuka-shi
Kanagawa-ken 259-12, Japan

Dr. Hiroichi Fujita — Phone: (044) 63-1141
Department of Engineering
Keio Gijuku University
832 Hiyoshi-cho, Kohoku-ku
Yokohama 223,Japan

PUBLICATIONS BY MEMBERS OF THE EDUCATIONAL TECHNOLOGY GROUP OF THE INSTITUTE OF ELECTRONICS AND COMMUNICATION ENGINEERS IN JAPAN (IECE)

I Papers in English

[1] Kurata, M. and T. Sato.
Test construction system applying item statistics. Proceedings of the International Conference on Cybernetics and Society, 1978, 368-372.

[2] Sato, T.
Instructional data processing, approach to computer managed instruction. NEC Research & Development, 1973, 29, 38-49.

[3] Sato, T. and M. Kurata.
Basic S-P score table characteristics. NEC Research & Development, 1977, 47, 64-71.

II Books in Japanese

[4] Sato, T. S-P table analysis: analysis and interpretation of test scores. Tokyo: Meiji-Tosho Publishing Co., 1975.

[5] Sato, T. (Ed.). CMI system (computer managed instruction system): Uses of computers in education. Denshi Tsushin Gakkai (Institute of Electronics and Communication Engineers of Japan), 1976.

III Other Publications in Japanese

[6] Fujiwara, H. A study on partitioning of S-P tables for learning diagnosis. Kodo Keiryogaku, (Japan Behaviometrics), 1979, 6, 1-9.

[7] Kurata, M. and T. Sato. Simulation and analysis of the item score table using an S-P score table model. Kodo Keiryogaku, (Japan Behaviometrics), 1976, 4, 11-17.

[8] Nagaoka, K. and H. Fujita. Analysis for speaking time in discussion. Kodo Keiryogaku (Japan Behaviometrics), 1978, 6, 1-8.

[9] Sato, T. Engineering techniques in the analysis of educational data V: application of entropy. NEC Central Laboratories, 1977.

[10] Sato, T. Construction of a test and the S-P table. NEC Central Laboratories, 1978.

[11] Sato, T. Hierarchical display of the network of teaching elements using the Interpretive Structural Modeling technique. Denshi Tsushin Gakkai (IECE), Educational Technology, 1978, 4, 23-28.

[12] Sato, T. How to make and use S-P tables as a method of analyzing the test result. Shido to Hyoka (Guidance and Evaluation), 1978, 282, 44-51.

[13] Sato, T. Item bank system: A set of test items and its cooperative use. Paper presented at the Third Conference on Educational Technologies, Naruto, Tokushima, 1978.

[14] Sato, T., M. Kurata and H. Ikeda. Estimation of statistical characteristics of the educational tests. Denshi Tsushin Gakkai (IECE), Educational Technology, 1978, 2, 27-30.

[15] Sato, T. and H. Chimura. Determination of hierarchical structure of instructional units using the Interpretive Structural Modeling method. Denshi Tsushin Gakkai (IECE), Educational Technology, 1979, 1, 11-16.

[16] Takeya, M. and T. Sato. On the learning progress distribution of programmed instruction. Kodo Keiryogaku (Japan Behaviometrics), 1975, 3, 12-21.

[17] Takeya, M. A property analysis of an item score matrix in CMI systems. Trans. IECE, 1977, 60, 967-974.

[18] Takeya, M. Hierarchical structure analysis among instructional objectives on student performance scores. Denshi Tsushin Gakkai (IECE), Educational Technology, 1978, 7, 23-28.

[19] Takeya, M. Application of an item structure analysis to an S-P score table. Denshi Tsushin Gakkai (IECE), Educational Technology, 1978, 12, 35-40.

[20] Takeya, M. On an expanded item relational structure analysis and its application. Denshi Tsushin Gakkai (IECE), Educational Technology, 1979, 1, 23-26.

[21] Takeya, M. Comparison of the item relational structure analysis based on item orderliness with other methods. Proceedings of the Conference of Nippon Kodo Keiryo Gakkai (Japanese Behaviometric Society), 1979, 102-105.

[22] Takeya, M. Item relational structure analysis based on performance scores for educational evaluation. Trans. IECE, 1979, 62, 451-458.

AD-A087 127 TENNESSEE UNIV KNOXVILLE DEPT OF PSYCHOLOGY F/G 5/9
RESEARCH ON THE MULTIPLE-CHOICE TEST ITEM IN JAPAN: TOWARD THE --ETC(U)
APR 80 F SAMEJIMA N00014-77-C-0360
UNCLASSIFIED ONRT-M3 NL

2 of 2
AD A
087127

END
DATE
FILMED
9-80
DTIC

[23] Takeya, M. Application of an item relational structure graph to internal structure analysis of tests. Trans. IECE, 1979, in press.

[24] Yamashita, H., H. Maejima, J. Yokoi, and M. Takeya. Structure analysis of instructural programs using item relational structure analysis I. Denshi Tsushin Gakkai (IECE), Educational Technology, 1979, 1, 27-31.

☆ U. S. GOVERNMENT PRINTING OFFICE : 1980 625-641/1973